

UNIVERSITY OF SOUTHAMPTON

FACULTY OF PHYSICAL AND APPLIED SCIENCES

Electronics and Computer Science

Using Linked Data in Purposive Social Networks

by

Priyanka Singh

Thesis for the degree of Doctor of Philosophy

September 2016

UNIVERSITY OF SOUTHAMPTON

ABSTRACT

FACULTY OF PHYSICAL AND APPLIED SCIENCES

Electronics and Computer Science

Doctor of Philosophy

USING LINKED DATA IN PURPOSEFUL SOCIAL NETWORKS

by Priyanka Singh

The Web has provided a platform for people to collaborate by using collective intelligence. Messaging boards, Q&A forums are some examples where people broadcast their issues and other people provide solutions. Such communities are defined as a Purposeful Social Network (PSN) in this thesis. PSN is a community where people with similar interest and varied expertise come together, use collective intelligence to solve common problems in the community and build tools for common purpose. Usually, Q&A forums are closed or semi-open. The data are controlled by the websites. Difficulties in the search and discovery of information is an issue. People searching for answers or experts in a website can only see results from its own network, while losing a whole community of experts in other websites. Another issue in Q&A forums is not getting any response from the community. There is a long tail of questions that get no answer.

The thesis introduces the Suman system that utilises Semantic Web (SW) and Linked Data technologies to solve above challenges. SW technologies are used to structure the community data so it can be decentralized and used across platforms. Linked Data helps to find related information about linked resources. The Suman system uses available tools to solve name entity disambiguation problem and add semantics to the PSN data. It uses a novel combination of semantic keyword search with traditional text search techniques to find similar questions with answers for unanswered questions to expand the query term with added semantics and uses crowdsourced data to rank the results. Furthermore, the Suman system also recommends experts who can answer those questions. This helps to narrow down the long tail of unanswered questions in such communities.

The Suman system is designed using the Design Science methodology and evaluated by users in two experiments. The results were statistically analysed to show that the keywords generated by the Suman system were rated higher than the original keywords from the websites. It also showed that the participants agreed with the algorithm rating for answers provided by the Suman system. StackOverflow and Reddit are used as an example of PSN and to build an application as a proof of concept of the Suman system.

Contents

Declaration of Authorship	xv
Acknowledgements	xvii
Nomenclature	xix
1 Introduction	1
1.1 Overview	1
1.2 Research Questions	3
1.3 Research Methodology	4
1.3.1 Scientific Steps to Answer KQ1 and KQ2	5
1.3.1.1 Knowledge Problem Investigation	5
1.3.1.2 Research Design	5
1.3.1.3 Research Design Validation	5
1.3.1.4 Research Execution	6
1.3.1.5 Analysis of Results	6
1.3.2 Scientific Steps to Answer DP1 and DP2	6
1.3.2.1 Problem Investigation	6
1.3.2.2 Treatment Design	7
1.3.2.3 Design Validation	8
1.3.2.4 Treatment Implementation	8
1.3.2.5 Implementation Evaluation	9
1.3.3 Scientific Steps to Answer KQ3, KQ4 and KQ5	9
1.3.3.1 Knowledge Problem Investigation	9
1.3.3.2 Research Design	10
1.3.3.3 Research Design Validation	11
1.3.3.4 Research Execution	12
1.3.3.5 Analysis of Results	12
1.4 Research Contribution	12
1.5 Structure of Thesis	14
2 Background	15
2.1 Emergence of Social Web	16
2.1.1 Web 2.0 and Social Media	17
2.1.2 Content-specific Social Networking Services	18
2.2 Collective Intelligence and Crowdsourcing	19
2.2.1 Information Sharing	20
2.2.2 Human Computation	21

2.2.3	Quality Management and User Recommendation	22
2.3	Semantic Web and Linked Data	23
2.3.1	Semantic Web Technology	23
2.3.2	Linked Data Technology	25
2.3.2.1	Benefits of Linked Data	26
2.3.2.2	Linking the Datasets	27
2.3.2.3	Linked Data Cloud	30
2.3.3	Semantic Web Vocabularies	31
2.3.3.1	FOAF	31
2.3.3.2	SIOC	32
2.3.4	Social Semantic Applications	33
2.3.4.1	Semantic Tagging	33
2.3.4.2	Semantic blogging and microblogging	34
2.3.4.3	Semantic Wiki	35
2.3.5	Semantic Search and Query	35
2.3.5.1	Information Retrieval	36
2.3.5.2	Web Search	36
2.3.5.3	Semantic Search	37
2.3.5.3.1	Keyword Disambiguation:	40
2.3.5.3.2	Concept mapping	41
2.3.5.3.3	Semantic Queries	42
2.3.5.3.4	Expert Recommendation:	44
3	Purposive Social Network	47
3.1	What is Purposive Social Network	48
3.2	Different types of communities in Purposive Social Network	49
3.2.1	Information Based Community	49
3.2.2	Interest Based Community	49
3.2.3	Expert Based Community	50
3.2.4	Location Based Community	50
3.3	Properties of Purposive Social Network	51
3.3.1	Community Size	51
3.3.2	Focused Interest	51
3.3.3	Direct Communication	51
3.3.4	Active Participation	52
3.3.5	Short Lifespan	52
3.3.6	Strong Incentive	52
3.4	Benefits of Purposive Social Network	52
3.4.1	Information Exchange and Self-interest	53
3.4.2	Symbiotic Relation and Social Exchange	54
3.4.3	Social Recognition and Personal Satisfaction	54
3.4.4	Recommendation System	55
3.4.5	Expert Finder	55
3.5	Challenges in Purposive Social Network	55
3.5.1	Recruiting and Retaining Users	56
3.5.2	Incentive Model	57
3.5.3	Quality Control	58

3.5.4	Search and Discovery of Quality Content	59
3.6	Case Study of Purposive Social Network	59
3.6.1	StackOverflow Analysis	60
3.6.1.1	Question and Answers	61
3.6.1.2	Users	62
3.6.1.3	Tags	63
3.6.1.4	Votes and Reputation	65
3.6.1.5	Communication network structure	66
3.6.1.6	Tags network	67
3.6.1.7	Incentive Design	68
3.6.1.8	Quality Control	71
3.6.1.9	Community Moderation	72
3.6.2	Reddit Analysis	72
3.6.2.1	Communication network structure	75
3.6.2.2	Subreddit network	76
3.6.2.3	Incentive Design	76
3.6.2.4	Quality Control	79
3.6.2.5	Community Moderation	79
3.6.2.6	Discourse Analysis	80
4	The Suman System	85
4.1	Use of Semantic Web and Linked Data in Purposive Social Network	85
4.1.1	Research problem and challenges	86
4.1.2	Benefit of Using Semantic Web and Linked Data in Purposive Social Network	87
4.1.2.1	Structured Data	87
4.1.2.2	Linking People to People and People to Data	88
4.1.2.3	Multidimensional Network and Graph	88
4.1.2.4	Integrated Knowledge	88
4.1.2.5	Smart Query and Search	88
4.1.2.6	Social Network Analysis	89
4.1.3	Research Validation	89
4.1.4	Research execution	89
4.1.5	Result evaluation	90
4.2	The Suman System	91
4.2.1	Problem Investigation	92
4.2.2	Treatment Design of the Suman System	92
4.2.2.1	Data Collection	93
4.2.2.2	Data Structuring	94
4.2.2.3	Keyword Annotation and Linking	94
4.2.2.4	Database and Query	96
4.2.2.4.1	Database Indexing and Configuration:	97
4.2.2.5	Suman Search Algorithm	98
4.2.2.5.1	Detailed Explanation of Each Step	99
4.2.2.6	Expert Finder	102
4.2.2.6.1	Detailed Explanation of Each Step	103
4.2.2.7	Design Innovation in the Suman System	105

4.2.2.7.1	Special feature of the Suman algorithms:	105
4.2.3	Design Validation	106
4.2.3.1	Validating the Suman system design	106
4.2.4	Treatment Implementation using StackOverflow and Reddit Datasets	107
4.2.4.1	StackOverflow Data Mining	107
4.2.4.2	Reddit Data Mining	108
4.2.4.3	Data Structuring using RDF	109
4.2.4.3.1	Generating RDF	110
4.2.4.3.2	Database	113
4.2.4.4	Keyword Annotation and Linking	114
4.2.4.4.1	Wikipedia Miner:	115
4.2.4.4.2	OpenCalais:	115
4.2.4.5	Information Retrieval	117
4.2.5	Implementation Evaluation	117
4.2.5.1	Keyword Annotation and Categories Analysis	118
4.2.5.2	Answers and Experts Analysis	119
4.2.5.3	Expert Recommendation Analysis	121
4.2.5.4	Linked Data Graph Analysis	122
5	The Suman System Evaluation	125
5.1	Knowledge Problem Investigation	125
5.2	Research Design	126
5.2.1	Knowledge Question 3 (KQ3)	126
5.2.2	Knowledge Question 4 (KQ4)	127
5.2.3	Knowledge Question 5 (KQ5)	128
5.3	Research Design Validation	129
5.3.1	Knowledge Question 3 (KQ3)	129
5.3.2	Knowledge Question 4 (KQ4)	130
5.3.3	Knowledge Question 5 (KQ5)	130
5.4	Research Execution	131
5.4.1	Calculating Sample Size	132
5.4.2	Selecting Questions	134
5.4.3	Questionnaire Design	136
5.4.3.1	Keyword Evaluation	137
5.4.3.2	Answer Evaluation	138
5.5	Result Evaluation	140
5.5.1	Dataset Frequency Distribution	140
5.5.2	Keywords T-Test	141
5.5.3	Q&A Correlation Test	144
5.5.4	Summary of the results	147
6	Conclusions	149
6.1	Summary of Research	149
6.1.1	Limitation of current systems	149
6.1.2	Research Questions	150
6.1.3	Research Contribution	151
6.2	Limitation of the Suman System	153

6.3 Future Work	154
A Experiment Questionnaire	155
A.1 Keywords Experiment	155
A.2 Answers Experiment	183
References	237

List of Figures

2.1	A taxonomy for collaboration alternatives (Albors et al., 2008).	17
2.2	Semantic Web Layer Cake Model (Krauss, 2014).	24
2.3	The hash URI and 303 URI approach with content negotiation (Sauer- mann and Cyganiak, 2008).	28
2.4	An RDF Triple.	28
2.5	Linked Data Cloud (Schmachtenberg et al., 2014).	30
2.6	Creating a FOAF profile and SIOC file of a user.	32
3.1	StackOverflow: Questions and answers posted every month	61
3.2	StackOverflow: Histogram of questions and answers posted the day of the week respectively	61
3.3	StackOverflow: User reputation histogram	62
3.4	StackOverflow: Tag trends per week of 5 most popular tags	63
3.5	StackOverflow network structure	66
3.6	StackOverflow: Popular tags clustered together	68
3.7	StackOverflow: Related tags clustered together	69
3.8	StackOverflow: Time histogram of questions receiving first answer and accepted answer respectively	71
3.9	StackOverflow: Questions and Answers Vote count histogram	72
3.10	Reddit: Posts created per month	73
3.11	Reddit: Histogram of posts created day of the week	75
3.12	Reddit network structure	76
3.13	Reddit users' link and comment karma histogram	78
3.14	Reddit: Time histogram of posts receiving first comment	78
3.15	Reddit: Posts and Comments Vote count histogram	79
3.16	An example of Reddit post (programmer564698, 2015) where people share personal stories	80
3.17	Example of funny StackOverflow response (Guy, 2008)	81
3.18	Example of funny StackOverflow response (Jeff, 2012)	81
3.19	Example of StackOverflow question (Skeet, 2009) where people are shar- ing their opinion	82
3.20	Example of StackOverflow question (Harvey, 2011) where people are shar- ing trivia	82
3.21	Example of Reddit post(carmichael561, 2013) where people are sharing pop culture references	83
4.1	The Suman system design components.	94
4.2	ER diagram of StackOverflow dataset	108

4.3	ER diagram of Reddit dataset	109
4.4	RDF schema of User's profile	110
4.5	RDF schema of StackOverflow posts	111
4.6	RDF schema of Reddit posts	113
4.7	Showing parent-child relationships in 2 different types of posts in Stack- Overflow and Reddit.	114
4.8	Wikipedia Miner annotation example	115
4.9	An example to keywords N-Triple linked to DBpedia	123
5.1	Calculating keyword experiment sample size using GPower	132
5.2	Calculating Q&A experiment sample size using GPower	133
5.3	Calculating participants sample size using GPower	134
5.4	Sample keyword experiment question	138
5.5	Sample Q&A experiment question	139
5.6	Keywords test frequency distribution diagram	140
5.7	Q&A test frequency distribution diagram	141
5.8	Keywords T-Test	142
5.9	Keyword T-Test showing confidence interval	142
5.10	Answers correlation test	144
5.11	Q&A data scatter plot diagram.	145
5.12	Q&A data scatter plot diagram showing questions' difficulty.	146

List of Tables

3.1	StackOverflow at glance as of June 2014	60
3.2	StackOverflow: Ten most popular tags and its instances	63
3.3	StackOverflow: Percentage of unanswered questions with number of tags .	64
3.4	StackOverflow: Difference percentage of unanswered questions in general tags vs specialized tags	65
3.5	StackOverflow: Percentage of Answers' vote count in general tags vs spe- cialized tags	65
3.6	StackOverflow: Number of users with reputations	70
3.7	StackOverflow: Questions and answers vote count	71
3.8	General overview of Reddit dataset	73
3.9	Reddit: Different types of posts	74
3.10	Number of subscribed users for each tag and subreddit in StackOverflow and Reddit respectively	77
3.11	Reddit: Number of users with karma	77
3.12	Reddit: Posts and comments vote count	79
4.1	StackOverflow: Percentage of questions with answers with confidence score	119
4.2	StackOverflow: Percentage of experts with confidence score	120
4.3	Reddit: Percentage of posts with answers and their confidence score . . .	120
4.4	Reddit: Percentage of experts with confidence score	120
4.5	Top users of top tags in StackOverflow	121
4.6	Top users of top disambiguated topics in Suman system.	122

Declaration of Authorship

I, Priyanka Singh , declare that the thesis entitled *Using Linked Data in Purposive Social Networks* and the work presented in the thesis are both my own, and have been generated by me as the result of my own original research. I confirm that:

- this work was done wholly or mainly while in candidature for a research degree at this University;
- where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated;
- where I have consulted the published work of others, this is always clearly attributed;
- where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work;
- I have acknowledged all main sources of help;
- where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself;
- parts of this work have been published as: (Singh and Shadbolt, 2013)

Signed:.....

Date:.....

Acknowledgements

This PhD, has been a hard and long process, but I wouldn't be here without the four pillars of support who helped me through everything.

The first pillar consists of all my supervisors and advisers who have helped me through my academic process. Especially my supervisors Prof. Sir Nigel Shadbolt and Prof. Elena Simperl. I am sincerely grateful from the bottom of my heart for helping me reach the end of my journey. I wouldn't have progressed without the help of my other two former supervisors, Dr. Manuel Salvadores and Dr. Gianluca Correndo. You have my gratitude for helping me along the way. I also want to thank Prof. Les Carr for his advice as my examiner and Lester Gilbert for his advice on my experiment design and analysis. All of you have guided me through my journey to finish my PhD. I wouldn't be here without you.

The second pillar is my family. I wouldn't have made it through my rough times without your unconditional love, comfort and strength. Maa, you have been my biggest ally. This thesis is dedicated to you and named after you. My sisters, Ankita and Arpita, you have been a shining beacon through the harshest storms, guiding me through rough patches. I am here because of your constant encouragement. I can't forget you Shivam, you have always assured me and reminded me how good things will be in the end. Most of all, I want to thank my father. Papa, I am here in this University, pursuing my higher studies, all because of you. Honorary mention to all my uncles, aunts, grandparent and cousins. Yash, now you can stop asking me when I will be done with my PhD.

The third pillar is all my friends who have been my family and made this place feel like home away from home. You have shared your stories, trials and turbulence. Because of you I never felt lonely in my journey. You made me feel I can do this and there is light at the end of the tunnel. Kewalin, Asim, Moody, Sumit, Eamonn, Jamal, Will, Nawar, Huw, Rikki, Jonny, Areeb, Alaa, Devasena, thank you all. Especially, Eamonn for proofreading my thesis. I also want to thank everyone in WAIS for your constant support, you have made me feel part of the family. My other friends who I have not mentioned here, I will always remember you and your support and encouragement.

The final pillar of support is all the people that I never met, but you have helped me out with my problems. This includes all the people in online forums who have helped me

fix my bugs, giving me advice on how to tackle any issues with my software and many other things that went wrong. I was able to fix my software and make it work because of your timely advice. This list also includes all the friends I made online that gave me a patient ear when I wanted to talk, the authors of all the books who have helped me stay sane and kept me distracted when I needed to forget about my work. Your help will always be remembered and I will always be grateful for your kindness.

Finally, Thomas, Eva, Chase, Rhea and Rick. You were always there with me when I needed you. You were always there when I didn't need you. It's a huge comfort knowing you will always be there.

Nomenclature

DP	Design Problem
FOAF	Friend Of A Friend
HTTP	Hyper Text Transfer Protocol
IDE	Integrated Development Environment
KQ	Knowledge Question
ODBA	Ontology Based Data Access
OWL	Web Ontology Language
PSN	Purposive Social Network
RDFa	Resource Description Framework-in-attributes
RDF	Resource Description Framework
RDFS	RDF Schema
RIF	Rule Interchange Format
RQ	Research Question
SIOC	Semantically Interlinked Online Community
SNS	Social Networking Services
SPARQL	Protocol and RDF Query Language
SSL	Secure Socket Language
WWW	World Wide Web
XML	Extensible Markup Language

Chapter 1

Introduction

1.1 Overview

This thesis discusses the concept of Purposive Social Network (PSN), a social network with a purpose, where people come together with a common objective to get information, build knowledge base, solve problems or achieve some common goal. Here people with similar interests and varied expertise may come together, use crowdsourcing techniques to solve a common problem and build tools for common purposes.

Web 2.0 has provided a distributed platform for people to collaborate to solve common problems. It has provided a platform to create Social Networking Services (SNS) for people to communicate and share information. In these social networks people are connected with each other using explicit relationships (friendships) or through data. This facilitates formation of PSNs. PSN are a type of SNS but not all SNS are PSN. PSNs are discussed in more detail in chapter 3.

People form groups and communities based on similar interests and purposes (McPherson et al., 2001). In many cases this may present opportunities to leverage collective intelligence for distributed problem solving and create knowledge. This crowdsourcing technique is different from human-based computing (Albors et al., 2008). Messaging boards, question-answer forums and wikis are examples of these types of social communities where people come together to create an emerging knowledge (Stenmark, 2002). In these websites people broadcast any problems they have and other users and experts answer and submit solutions.

Nowadays, there are many different SNSs created for different purposes and they target different usergroups. Different websites provide different functionalities. For example, many people use Facebook ¹ to connect with friends and Twitter ² for microblogging.

¹<http://www.facebook.com/>

²<http://twitter.com/>

There are also content specializing services, for example YouTube ³ for videos, Flickr ⁴ for pictures, etc. Many of these websites are centralized, users sign up, create a profile, invite their friends and create friendships. These websites can be open, meaning that all the data is visible to all the users, or closed, where only logged-in users can view any information. In most cases, the websites are semi-open, i.e. registered users can see everything and non-registered users can see partial information on the website. Hence, all the information is not accessible to everyone.

The main drawback of current PSNs is that they are all closed or partially open networks and the service providers own all the users' data, i.e. users cannot take their data away with them once they delete their account and leave that website or the website closes down. If people want to sign up to different websites to use their functions, they have to undergo the same cycle of creating their user profile and adding friends to their network again. They cannot migrate their data and social network from one website to another. All their data, activities and relationships are trapped inside a silo. In many countries, users of PSNs do not own the data they generate. The data are controlled by the websites. In some websites, if any user wishes to delete their account they lose all their data. They cannot export their data or their interaction with their friends and cannot migrate to another website. These PSNs are centralized and have their own APIs and structure. People cannot easily work across platforms and share information across different social networks. Fitzpatrick referred to this whole cycle of recreating profiles and reconnecting with friends across multiple SNSs as Social Network Fatigue (Fitzpatrick and Recordon, 2007).

Difficulties in the search and discovery of information in these PSNs are another issue. A given search engine can only retrieve information when explicitly asked. It does not return the solution of a problem if the solution doesn't exist on a webpage. Search engines only use the keywords in the search query to retrieve results. They might not expand the query to broader or narrower categories. Also, most search engines typically cannot access closed communities and websites. The search function of an individual website will only return results from their own network. People who want the advice and opinion of an expert can join a forum or messaging board or join an email list, but they can only find the subset of experts registered for that service, while losing a whole community of experts in other networks or communities.

Another issue users face in PSNs is not getting any response from the community on their post and then they have to go to different PSN to find the solution to their problem. For example, StackOverflow ⁵ is a technology based question and answering forums. StackOverflow have more than 3.2 million questions and 1.2 million registered users. Despite so many users, 23.7% of questions in StackOverflow do not get any answers

³<http://www.youtube.com/>

⁴<http://www.flickr.com/>

⁵<http://stackoverflow.com/>

(Singh and Shadbolt, 2013). There is a long tail of questions that get no answer, response or votes. This is a problem in many forums and question answering websites. These websites follow the power law of the Web (Albert et al., 1999). There are some popular posts with most response, votes and comments and there are many posts in the long tail that have very few or no replies at all. This is a common problem in many PSNs. It is discussed in more details in Chapter 3.

1.2 Research Questions

The main research question this thesis focus on is to investigate whether Linked Data and Semantic Web technologies can help to answer unanswered questions on PSNs (RQ1).

This research question (RQ1) is broad and it can be further broken down into design problems and knowledge questions as done in design science by (Wieringa, 2014). Design problem is about designing an artifact (a system or algorithm) to solve a research problem, improve something for stakeholders or investigate the performance of the artifacts and knowledge questions study an artifact in context.

The first step to answer RQ1 is to define PSNs (KQ1). This is a knowledge question and existing systems can be studied in the context of PSNs to define the main properties and characteristics of PSNs. Furthermore, other aspects of PSNs could be studied to see how PSNs are formed and the communities sustain themselves. This leads to study of community formation, community sustenance and main motives and incentives of users to join the PSNs to solve problems (KQ2). It is important to know how the community was formed as well as how it is maintained. So, these aspects of PSNs are studied in this thesis and these questions are answered by studying StackOverflow and Reddit⁶. These websites could be studied to see how it is structured and how they motivate and incentivize users to participate.

The RQ1 can be further broken down into design problems. A prototype system can be built or a software based solution could be used to investigate if Linked Data and Semantic Web technologies can provide solutions to the unanswered question in PSNs. This design problem is broken down into smaller problems. Firstly, it is investigated if Semantic Web and Linked Data technologies can be useful in search and discovery of answers in PSNs (KQ3). Secondly, a prototype system should be built that takes questions, answers, user profiles and other crowdsourced data (votes, favourites, tags, etc.) from PSNs and use these data to provide answers to unanswered questions (DP1).

For this thesis, a prototype system called the Suman System is built to answer the RQ1 and to solve the design problems. Moreover, the other problem this thesis aims to answer is if the community and users' data from PSNs could be used to find experts in a topic

⁶<http://reddit.com/>

and recommend them to users to answer unanswered questions (DP2). StackOverflow and Reddit websites are used to investigate this design problem. After creating the prototype, the Suman system needs to be evaluated to check if it meets the design goals and answers RQ1. This leads to another knowledge question that finds out how well the Suman system works and answers the unanswered questions in PSNs (KQ4) and how good are the recommended experts and if they could answer the unanswered questions or not (KQ5).

In summary, this thesis aims to investigate if Semantic Web and Linked Data technologies improve search and discovery of answers and experts in PSNs. StackOverflow and Reddit datasets are also analyzed to show how PSNs are formed and users participate. Finally, the Suman system is created as a proof of concept that helps to answer these knowledge questions and design problems and it is evaluated to show it answers the research questions.

1.3 Research Methodology

The aim of this thesis is to investigate if Semantic Web and Linked Data can be used to create a PSN where data is open, linked and have added semantics (knowledge). Furthermore, the added links and semantics can help search and discovery of information. This would theoretically help solve the problems discussed above. This is done by first studying PSNs and to use the knowledge gained from the study to improve search and discovery of information in the Suman system.

Design science research methodologies for information systems and software engineering research can be used as a guideline to perform this research and answer the research question. (Wieringa, 2014) breaks down design science research methodologies into two activities: Design problem and Knowledge question. This thesis and research will follow these methodologies to describe the research done to answer the RQ1.

(Wieringa, 2014) describes the design problem as designing an artifact that improves something for stakeholders and empirically investigating the performance of an artifact in the context. The artifacts could be methods, algorithms or frameworks used in software and information systems. The context of these artifacts is the design, development and use of software and information system. The artifacts are designed for these contexts and it should be investigated in those contexts. This forms the knowledge question. Both the design problem and empirical research can be treated as a problem-solving approach. These approaches can be described using engineering cycle and empirical cycle scientific steps respectively. In the first, the design artifacts intend to help the stakeholders. In the second, the knowledge question about the artifacts are answered in the given contexts.

The results of both engineering cycle and empirical cycle are fallible. The designed system might not fully meet the goals of the stakeholders and answers to the knowledge questions might have limited validity. So, the designed system and answers produced must be justified. This is done by validation of the designed system in terms of stakeholders' goals and requirements and the validation of inferences in the empirical cycle. These scientific steps that follow these methodologies are discussed below.

1.3.1 Scientific Steps to Answer KQ1 and KQ2

The RQ1 is about finding answers for unanswered questions in PSNs. In this thesis, the term Purposive Social Network (PSN) is used to describe systems where people with similar interest come together and solve each others problems using crowdsourcing techniques. So PSNs needs to be defined and study. KQ1 and KQ2 is all about understanding and studying PSNs. The details of PSNs and all the scientific steps taken to do this is described in Chapter 3. This section gives a brief summary of how KQ1 and KQ2 are answered.

1.3.1.1 Knowledge Problem Investigation

KQ1 is to define PSN. This thesis is about PSNs, so understanding PSN is important. This is done in KQ2. Here the community structure of PSNs, users' behaviour and motive and incentive models provided by PSNs are studied.

1.3.1.2 Research Design

To define PSNs, literature in the area of social network and community formation is critically analyzed. Existing PSNs are studied and their main characteristics and properties are defined using the literature. This is a theoretical question and it is answered using the research done in the area and existing examples of PSNs. This answer KQ1.

To study PSNs, StackOverflow and Reddit websites are used. These popular websites are used by software developers and the data is open. All the user data are analysed statistically to understand PSNs. The community and social network data are analysed to understand the community structure and how users are linked to each other. This answer KQ2.

1.3.1.3 Research Design Validation

There are a lot of research done in the area of social networks and social network analysis as seen in Chapter 2 and 3. The literature and current examples of PSNs provide enough

support to define PSNs and understand its characteristics and properties. This answered KQ1 but the answer is subjective and fully answered.

The statistical tests done to analyse StackOverflow and Reddit data are standard tests. They are replicable and open for people to test. The analysis done is valid and justified. The KQ2 is fully answered, but there are many other statistical tests that could be performed to get more information about the PSNs.

1.3.1.4 Research Execution

StackOverflow and Reddit data are analysed in details in Chapter 3. Simple statistical analysis is done on user data and behaviour. Social network analysis is done on community data to find the community structure.

1.3.1.5 Analysis of Results

The main findings of StackOverflow and Reddit data analysis can be found in Chapter 3. This shows how popular some tags and subreddits are and how strong some communities are. The smaller communities are also active, but they dissipate easily.

The study of existing literature and PSNs answer KQ1 and study of StackOverflow and Reddit answers KQ2.

1.3.2 Scientific Steps to Answer DP1 and DP2

The RQ1 is about solving a problem in PSNs, searching for answers to unanswered questions in PSNs using Linked Data and Semantic Web technologies. This is a design problem and that is solved by the Suman system. The details of the design problem and the scientific steps taken to solve this problem is discussed in Chapter 4. A short summary of the research methodology is given below.

1.3.2.1 Problem Investigation

RQ1 is about finding answers to unanswered questions in PSNs. A solution based approach is required to solve this problem. A system or search algorithm should be designed to provide a solution to this problem (DP1). Another approach to solve this problem is to recommend experts that could potentially answer this unanswered question. This expert recommender system is another solution to the problem (DP2).

Building this system using Linked Data and Semantic Web Technologies arises some design challenges. They are :

1. Data Collection and integration: The main challenge here is to collect and integrate the structured data from different PSNs so it could be used by the search algorithm.
2. Name Entity Disambiguation: The main challenge here is to identify the main topic and concepts of the questions and answers to the added semantics could be used by the search algorithm.
3. Semantic search and query: The main challenge here is to use all the PSNs datasets that are semantically enriched and linked and use it to improve the search for answers to the unanswered questions.

These 3 challenges need to be met in the engineering cycle of the Suman system.

1.3.2.2 Treatment Design

The Suman system is created as a proof of concept to answer DP1 and DP2. The main goals of the research and the Suman system are searching for answers to unanswered questions in PSNs and to search for experts that could possibly answer the unanswered questions in the PSNs.

Suman is a Sanskrit word meaning wise and good mind. The Suman system creates a distributed PSN that links the data across different websites. The Suman system uses Semantic Web technologies to integrate heterogeneous datasets from PSNs to solve the data silo problem. The Semantic Web, as envisioned by Sir Tim Berners-Lee, is an extension of the WWW that enables people to share content beyond websites, applications and platforms (Berners-Lee et al., 2001). It is an intelligent web of structured data where each resource has a URI and is represented in RDF triples (Klyne and Carroll, 2006). Semantic Web technologies provide tools to represent and structure social network data to make it portable. This makes it possible to connect data from different portals to form a decentralized system that can be easily queried across platforms and reused. The Solid project is a step in this direction (Conner-Simons, 2015).

The Suman system also uses Linked Data and Semantic Web technologies for name disambiguation of keywords and topics. The structured data makes it machine-readable and ontologies helps to describe the data and give it meaning, understandable by both machines and humans (Berners-Lee et al., 2001). This data, when linked to other datasets and with the semantics added, can create a network of interlinked categories. It can be used to search for extra information and provide more details to the queries. This helps to solve the problem where users do not find any reply to their questions.

The Suman system disambiguates a topic and create a document and keyword graph. Further on, the Suman system uses the search algorithm that combines keywords based

semantic search with traditional text based search to find answers for unanswered questions from PSNs. The algorithms use SPARQL (Prud'Hommeaux et al., 2008) queries at SPARQL endpoint and uses the crowdsourced data (votes) to rank the results.

The same techniques can be used to create a semantic rich user profile and the experts can be linked to categories. The Suman system also searches for experts in the field and recommends them to provide answers to the queries with no results. This helps in improving search and discovery of answers for questions and experts in a field and provides useful information that is otherwise unavailable in the website. This helps in forming a PSN with right users and community. More details of Suman system can be found in Chapter 4.

1.3.2.3 Design Validation

The Suman system has specific requirements to search for answers to unanswered questions from PSNs using Linked Data and Semantic Web technologies. The system will be considered valid if it meet the above mentioned requirements.

The Suman system searches for answers to the questions from the preexisting knowledge-base from PSNs. It solves the all the 3 design problems mentioned earlier. It searches for answers thus meeting the design goal for DP1 and recommends experts meeting the design goals for DP2. It meets the goals and solves the problems for the stakeholders.

Although, DP1 goals are fully met, DP2 goals are partially met. The Suman system recommends the experts, but it is not certain that the experts will provide the answer the user requires.

1.3.2.4 Treatment Implementation

The Suman system is built using real world PSNs data. The system focuses on technology based question and answering forums like StackOverflow and Reddit for evaluation purposes. These websites are good example of PSN. In these websites users ask technical questions to the community and experts in the field provide solutions. These websites are quite popular among software programmers and developers to ask questions, share information and have discussions.

Data from these websites are collected and structured into RDF. Next, they were analysed using tools like Wikipedia Miner and OpenCalais to do Name Entity Disambiguation. The disambiguated topics and keywords were then linked into DBpedia and categories were added to create a keyword matrix. The search algorithm then used crowdsourced data to search and rank answers and experts. The details can be found in Chapter 4.

1.3.2.5 Implementation Evaluation

The Suman system was evaluated using StackOverflow July 2014 data. There were 20,326 unanswered questions in the top 10 tags and the system searched for relevant answers with confidence score more than 75% for 13,209 questions. 23.62% of unanswered questions had one or more answers with confidence score of 85% and 82.27% of unanswered question has a confidence score of more than 50%.

The search result also showed the list of experts recommended to answer each unanswered question. The list followed the same pattern as an answer. It only showed the top 5 experts with confidence score higher than 50%. If the question didn't have any answer with confidence score greater than 50% even then the system would show a list of recommended experts. This list was also restricted to the experts with confidence score of greater than 50%. More details can be found in Chapter 4.

The Suman system DP1 and DP2 were fully met as it answered the questions and recommended experts.

1.3.3 Scientific Steps to Answer KQ3, KQ4 and KQ5

The Suman system uses Linked Data and Semantic Web technologies to search for answers for unanswered questions. The system needs to be evaluated and the approach used the the Suman system needs to be justified. This is broken down into KQ3, KQ4 and KQ5. This is discussed in detail in Chapter 4 and 5. The brief summary is given below.

1.3.3.1 Knowledge Problem Investigation

KQ3 investigates if Linked Data and Semantic Web Technologies can be useful in search and discovery of answers for unanswered questions in PSNs. This question can be answered by showing Linked Data and Semantic Web technologies can help solve the research challenges discussed in section 1.3.2.1.

KQ4 investigates how well the Suman system answers the unanswered questions in PSNs.

KQ5 investigates how good are the recommended experts and if they can answer the unanswered questions in PSNs.

1.3.3.2 Research Design

KQ3 tests if Linked Data and Semantic Web technologies provide useful tools in search and discovery on answers. There are systems and algorithms that can search for information without using the Linked Data and Semantic Web technologies. The PSNs data used in this thesis are from StackOverflow and Reddit. These data is from the technology domain and have structured knowledge base. The existing ontology and knowledgebase on open Linked Data consists of many data sources from technology domain. Adding semantics in StackOverflow and Reddit data source makes it machine readable and adds knowledge. This helps to identify the meaning and core topics in the datasets. Furthermore, Linked Data links this topic to the other datasets, thus adding more information and helping to categorise and organize the datasets.

So, using Semantic Web and Linked Data technologies is a useful idea in the PSNs. This helps to understand the underlying meaning and concepts of the question and answers. This added semantic could potentially improve the search and discovery of information. The usefulness of Linked Data and Semantic Web technologies in search and discovery of information has been discussed in details in Chapter 2 and it is supported by the literature review.

Then the question arises, what is the best way to employ Semantic Web and Linked Data to the PSNs datasets. This could be done by doing a Name Entity Disambiguation to the questions and answers to understand the main topics and concepts of the question and answers. This adds semantics to the datasets. These topics could be linked to the Linked Data Cloud to help form connections between multiple datasets and use the relationship to infer more information. The added links provide additional knowledge and information that could potentially help to improve search and discovery of information.

The usefulness of the semantics and linking needs to be measured to justify using it. This can be done using a user experiment where participants and look at the added semantics and rate how well these keywords describe the question and answers. The other was to measure the usefulness of the added semantics is to analyse the relationship and categories of the topics in the questions and answers. These relationships and categories help to categorize and structure to the knowledge and help to find similar questions and answers. It is hypothesized that this helps in improving search and discover of answers to the unanswered questions. The usefulness of this could be measured by another user experiment where participants can rate how well the search result can answer the unanswered question. Statistical analysis could be done on the data collected from these experiments. If the finding and results and statistically significant, then it could be inferred that using Semantic Web and Linked Data improve the PSNs and answer the research questions. This would justify the use of these technologies and we could be satisfied with the end result because it meets the research goals.

Furthermore, the main advantage of using Semantic Web is to use the added meaning and relationship between data to improve the system. This added meaning and concepts can be tested by user experiment where users are asked to rate how well the concepts added by Semantic Web technologies describe PSN data. There are tags and categories to questions and answers in StackOverflow and Reddit data. This could be compared to added keywords from Semantic Web tools. A T-Test could be performed to see if this is useful or not.

KQ4 tests if the answers searched by the Suman system is good or not. This could be measured by a user experiments where users who have expertise in the field can check if the answers provided by the Suman system can answer the unanswered questions. Participants can rate the answers and this rating could be compared to the Suman search algorithm rating. The correlation between the two ratings could be measured to see how well the system searches for answers to unanswered questions.

KQ5 test is harder to design. This is because it recommends experts in the field. It doesn't show the answers. It is not possible to contact the recommended experts and ask them to answer the unanswered questions and then test their answers. It is also hard to show the complete user profile to participants in a user experiment and then rate the expertise of the recommended experts. Hence, in this these KQ5 is not tested and it is one of the limitation of the thesis.

1.3.3.3 Research Design Validation

The T-Test designed to measure the usefulness of the semantics added to PSN data using Linked Data and Semantic Web technologies is a standard statistical test. It is used in the scientific community to compare means of two pairs of the dependent variable. The standard measurement of the p value and r value justifies the experiment results. Furthermore, it will fully answer the KQ3.

The Correlation test designed to measure KQ4 to test how well the search result produced by the Suman system answers the unanswered questions is another standard statistical test used and accepted in the scientific community. The experiment design is justified and if the measurement of p and r value meets the scientific community requirement, then the test results are accepted and inference is justified. This experiment and the inference made by the results will fully meet the requirement to answer KQ4.

There is no suitable experiment designed to answer KQ5, hence this question is not answered at all and remains one of the drawbacks and limitations of this thesis.

1.3.3.4 Research Execution

The Suman system has been evaluated to make sure the added keywords are meaningful and the answers provide a solution to the questions by doing two user experiments. Both the experiments use standard sample size and follow the scientific steps of the experiment and analysis. It is discussed in detail in Chapter 5.

The first user experiment asked participants to rate the Suman system generated keywords compared to the original keywords to questions and answers. 30 questions and answers are tested by 20 participants. A T-Test is performed and the results are analysed. This experiment aims to answer the KQ3.

In the second user experiment, participants were shown an answer that could provide a solution to an unanswered question. They were asked to rate how well the answer answered the question. The participants rating and the Suman system algorithm rating were correlated. 46 questions and answers were tested by 20 participants. The result of this experiment aims to answer KQ4.

1.3.3.5 Analysis of Results

The results of both the experiments were statistically analysed. For the first experiment, it showed that the keywords generated by the Suman system were rated higher than the original keywords from StackOverflow and Reddit in 63.3% of the cases. It is inferred that the added keywords had some benefits to the PSN data and Linked Data and Semantic Web technology is useful in PSN. This answered KQ3.

The analysis of the results of the second user experiment showed that the participants agreed with the algorithm rating for answers provided by the Suman system. There was a positive correlation between the two ratings ($r = .380$, $n = 46$, p (two-tailed) = .009). The correlation between the two ratings was moderately strong and the significance was $<.01$. It is inferred that the Suman system did moderately well in answering the unanswered questions in PSN. This result answered KQ4.

More details of experiments and results is in chapter 5. But overall the Suman system performed well and showed that Semantic Web and Linked Data technologies can be used to provide answers to unanswered questions in PSNs.

1.4 Research Contribution

The primary aim of the research represented in this thesis is to show that Linked Data and Semantic Web technologies can be used to create an open and distributed system, and improve the search of answers and experts in a crowdsourced Q&A system.

To achieve this, first PSN is defined and studied in detail. Different attributes and characteristics of PSN are identified, how it is formed, why it is created is also studied. StackOverflow and Reddit fulfill the criteria of PSNs. The datasets are collected from these two websites and the emergent knowledge generated by the community is analyzed as a study of PSNs. These are done by following the design science methodology to answer knowledge questions (KQ1 and KQ2) and both the questions have been fully answered. Although it is noted that the KQ1 is a subjective question and can have many different answers. KQ2 is statistical analysis and it is noted that there are many different analysis could be done to find more details about PSNs.

The Suman system is created using the design science methodology. It solves the DP1 and DP2 questions discussed in the research question. The Suman system harvests data, cleans it and then structure it into RDF. The name entity disambiguation problem is solved using Wikipedia-Miner and OpenCalais. The keywords are categorized and mapped to the DBpedia and OpenCalais concepts. The Suman system also creates a weighted keyword graph for each question and answer. The algorithm searches for answers to unanswered questions and recommend experts in the field.

The search results provided by the Suman system are evaluated and analyzed. The Suman system is analyzed by conducting two experiments (more details in chapter 5). The first keywords experiments measure the usefulness of the system generated keywords. This is done by asking participants to rate how well the original keywords and the system generated keywords describe a question. The analysis of the experiments shows that the participants rated the Suman system generated keywords better than the original keywords in 63.3% of the cases. This answer KQ3.

The second experiment was to measure the answers provided by the Suman system to unanswered questions. The participants were shown an answer to an unanswered question. They were asked to rate how well the answer provided a solution to the unanswered question. The Pearson correlation test was performed to measure the relationship between the participants rating and Suman algorithm rating. There was a positive correlation between the two variables ($r = .380$, $n = 46$, p (two-tailed) = .009). This answer KQ4.

The results of the experiments showed that the Suman system provided answers to the unanswered questions and solved the problem that the thesis focus on. It helps to find answers to unanswered questions in PSNs using Semantic Web and Linked Data technologies.

1.5 Structure of Thesis

This thesis is structured into 6 chapters. This is the first chapter that introduces the research question, the research methodology and research contribution.

Chapter 2 gives a background and history of social networking and collective intelligence. It also describes the role of Semantic Web technologies in this area. It explains the evolution of the Semantic Web and Linked Data in the area of Social Networks briefly introducing some of the most important and widely used social semantic technologies and applications used nowadays. The chapter also discusses different semantic search and query techniques and related research in expert recommendation.

Chapter 3 defines and describes in more detail the topic of PSNs and community formation. It discusses the motivation of the research explaining why forming a quick and purposeful community is important and how Linked Data can help in achieving this goal. This chapter also analyses StackOverflow and Reddit as examples of PSN. The questions, answers and user information are analyzed to describe the network ties and relationships. The individual role of users is also studied and the incentive model of the website is discussed that motivates user's for their quality contribution. This answer KQ1 and KQ2.

Chapter 4 provides details of the design methodology used to design the Suman system and how the data was collected, cleaned and structured. It also discussed different tools used to solve the keyword disambiguation problem and how the keywords were linked to the Linked Data Cloud. This chapter also describes the search algorithm that the Suman system is used for concept mapping and creating the document keyword graph. It also mentions combining the keyword based semantic search with text search using the crowdsourced data to improve the search of answers and experts. Finally, this chapter evaluates the Suman system design.

Chapter 5 describes the experiment designed to test and evaluate the Suman system by users and how the data was collected. This is later analyzed by different statistical tests. This chapter shows how Semantic Web and Linked Data technologies can help in topic disambiguation, community formation, integration and search and discovery of better information and experts. This answer KQ3 and KQ4.

Chapter 6 describes potential directions for future work with the Suman system and closes with some concluding remarks on this thesis.

Chapter 2

Background

The Web has brought people from different parts of the world, socioeconomic status and culture together. It provides a platform for people to connect with friends, families, and strangers to share ideas and information. There are 3.2 billion people connected to the Internet as of June 2015 (ITU, 2015). It has become easier for people to communicate and collaborate with each other on the Web. People do not only use it to share information, they also use it to form groups and communities with like-minded people and create their own social network.

There are different types of websites and tools available for social collaboration. People can use them to create a collective knowledge base. Wikipedia is an example where people created knowledge together in a collaborative community. There are many AI-complete problems that are harder for computers to do, but people and the power of collective intelligence can easily solve it. Humans are better at natural language processing, speech recognition, and image-processing tasks. Human computation and crowdsourcing techniques can be used to solve these problems (von Ahn, 2009). The social Web and human computation offer many opportunities to solve difficult computational problems and create an emergent knowledge thus forming a Purposive Social Network (PSN). This is a new concept defined in this thesis and discussed in details in the next chapter.

This chapter discusses the emergence of the social Web and how the Web has evolved to Web 2.0 that connects all the objects and people. In recent years, Semantic Web technologies are also incorporated in the Web to create a new era of Web 3.0 (Hendler, 2009). These technologies could extend the scope of the current social Web and social media. The Semantic Web and Linked Data technologies provide tools to open and link the data, combine information and search them. This chapter also discusses how Linked Data links different datasets and how different semantic search and query techniques can improve information retrieval. This chapter provides literature support for the design decision made in the Suman system, there is no new and innovative work added

here to judge the proposed system. The literature reviewed and critically described in this chapter is to support why Linked Data and Semantic Web technologies can help improve the knowledge in the PSN. It also describes different semantic search techniques, algorithms and systems out there that inspired to build the Suman system.

2.1 Emergence of Social Web

The invention of the World Wide Web (WWW) and web-browsers allowed more people to interact with each other, share documents and create static webpages easily, without having a detailed technical and specialized knowledge. Although Tim Berners-Lee had envisioned a read-write Web, where the very first browser also worked as an editor (Berners-Lee et al., 2006b), the initial WWW was a read-only Web for the majority of people. In the early days, the Web was mostly a collection of webpages with a phone book like directory to look up individual websites that were connected using hyperlinks (Berners-Lee et al., 2000).

The passive attitude towards the Web changed when the web-browsers and search engines made webpages easily accessible to users. The business world adopted the technology and started using it for E-business. More communication tools like IRC and MS net meeting were developed for organizations to communicate and collaborate. Organizations started to have their personal intranet and portals as a collaborative tool and to integrate information (Stenmark, 2002).

The Web was adopted by more organizations, especially the news and entertainment industry and more and more information was put on the web as RSS (Board, 2007). People moved from a static, personal webpages to blogs and wikis and started the social transformation of the Web from web of data to web of people (Albors et al., 2008). Figure 2.1 categorizes different collaboration methods used by people based on how strongly they are connected with each other and how much information they share with each other.

This section discusses the concept of Social Web because the PSNs are part of the social web where people come together and form communities. The technologies and systems studied in this section give an insight on how social web forms and grows, this helps to understand how PSNs can form and grow. Furthermore, PSNs are social applications that have focused interest and these applications could be content specific. These literatures are discussed here because they share intrinsic properties with PSNs. The main properties of PSNs discussed in details in Chapter 3 but in this section it gives the insight about other social networking services.

tracked over 181 million blogs growing from 36 million in 2006 and there are 1 million new posts everyday (Nielsen and McKinsey, 2011).

Nowadays, there are blogs in every category and topic and blogging has turned into the means to share news and information instantly. The Blogs are densely interconnected based on political ideologies, language, topics and categories (Adamic and Glance, 2005; Chi et al., 2007). They have made it easier for users to create new content and share it on the Web. The simple editor does not require any technical knowledge and blogs provide a platform for people to create network with similar interest and expertise.

2.1.2 Content-specific Social Networking Services

In 2003, an SNS called Friendster ³ had attracted 5 million. Soon many other similar websites emerged, of which Facebook was the most successful with more than 1.31 billion active users as of June 2015 (Facebook, 2015). People use these SNS in every aspect of their life. They use it to share personal information with friends and family (Facebook) or they use it to share their work and interest with colleagues (LinkedIn). Business enterprises have also adopted SNSs to connect to their consumers and advertise the products directly to the people who are interested in it (Ridings and Gefen, 2004).

The SNS connect people together through personal connection as well as through shared objects. A good example of this is Facebook 'Like' button. This is used to connect people who like the same music or movies or share similar interest. This creates a vast community of people who share the same interest and also share similar information (Joinson, 2008).

There are also microblogging services like Twitter and Tumblr ⁴ that makes sharing information easier and faster. These services create implicit relationships between people by the follower-following method. The social network grows quickly as people do not need to approve any friendship and the content is distributed widely and rapidly by the re-sharing feature they have. Users can retweet the same tweet on Twitter and reblog the same posts in Tumblr multiple times creating a viral distribution of information (Huberman et al., 2008).

After Facebook, a new trend emerged for content specific social networking websites like YouTube to watch videos, Flickr to share pictures, Delicious ⁵ to share bookmarks, LastFM ⁶ and Spotify ⁷ to listen to music, etc. In these websites people connect with other people with similar interests. They form groups and communities and the websites

³www.friendster.com/

⁴<https://www.tumblr.com/>

⁵<http://delicious.com/>

⁶<http://www.last.fm/>

⁷<http://www.spotify.com/>

take advantage of users' ratings for quality control and users' behaviour for recommendation (Linden et al., 2003; Bennett and Lanning, 2007). These websites depend on users' contributions to sustain the community.

2.2 Collective Intelligence and Crowdsourcing

The Web provided a platform for people to share, discuss ideas and information, and collaborate with each other. The Web 2.0 main trend was using the collective intelligence of people to create knowledge. Web 2.0 used the wisdom of crowd to solve problems, perform quality control and provide recommendations. PSNs use crowdsourcing techniques to create their knowledgebase and solve problems. Understanding different types of crowdsourcing systems provide useful information and helps to design PSNs. The papers discussed in this section provide insight about different types of crowdsourcing systems and how they form and grow. These systems can be studied to design better PSNs.

Jeff Howe coined the term crowdsourcing in 2006 to describe how businesses outsource their work to the people using the Internet (Howe, 2006). (Estellés-Arolas and González-Ladrón-de Guevara, 2012) later studied different definition of the term and integrated it together into this definition. This is the definition adopted for the term crowdsourcing in this thesis.

"Crowdsourcing is a type of participative online activity in which an individual, an institution, a non-profit organization, or company proposes to a group of individuals of varying knowledge, heterogeneity, and number, via a flexible open call, the voluntary undertaking of a task. The undertaking of the task; of variable complexity and modularity, and; in which the crowd should participate, bringing their work, money, knowledge ****[and/or]**** experience, always entails mutual benefit. The user will receive the satisfaction of a given type of need, be it economic, social recognition, self-esteem, or the development of individual skills, while the crowdsourcer will obtain and utilize to their advantage that which the user has brought to the venture, whose form will depend on the type of activity undertaken." (Estellés-Arolas and González-Ladrón-de Guevara, 2012)

Crowdsourcing uses collective intelligence where a 'form of universally distributed intelligence, constantly enhanced, coordinated in real time, and resulting in the effective mobilization of skills' (Lévy, 1997).

There are many different types of crowdsourcing systems and they target different user base. There are some systems that use gamification techniques to get users' input (e.g. ESP Games (von Ahn and Dabbish, 2004)) and other systems pay the workers to finish

some tasks (e.g. MechanicalTurk (Buhrmester et al., 2011)). There are also community projects like Wikipedia ⁸ where people create knowledge for altruistic reasons.

A crowdsourcing system can be categorized into many different ways (Yuen et al., 2011). This thesis focuses on three aspects and application of such systems.

- Information Sharing
- Human Computation
- Quality Management and User Recommendation

2.2.1 Information Sharing

One of the popular applications of crowdsourcing systems is using the wisdom of the crowd to create knowledge and share information. Here are some case studies where people share information to solve problems.

Reddit is a popular website that allows users to share news, pictures, links, any kind of information and people vote popular items and have discussions on any topics. NoiseTube is a web app that made it easier for users to use their mobile phone to measure noise in an area, add GPS tags to it and share it with other people to monitor noise pollution (Maisonneuve et al., 2009).

One of the widely used crowdsourcing tools to create a collective knowledge is a Wiki. Web 2.0 provided easy to use tools for collaborative authoring and ownership of information. A wiki allows users to add, remove, edit or modify any content on a webpage and keep track of different version changes. Wikipedia, a collaborative encyclopedia, is an excellent example where people with similar interest and expertise came together to create knowledge. Crowdsourcing is used for quality control and to stop vandalism spamming. These collaborative networks strive because they enforce a strong sense of community with people through collaboration and creation of content where individual contribution was counted and it became an efficient knowledge management tool (Kane, 2009).

Question and answering systems are another example where people answer other people's questions and create knowledge. Yahoo! Answers ⁹ and StackOverflow are some of the widely used systems. There are also many messaging boards and forums where people discuss topics, ask questions and provide suggestions. Yahoo! Suggestion Board ¹⁰ is a feedback and suggestion system. It will be discussed in detail in chapter 3 but these kinds

⁸<http://www.wikipedia.org/>

⁹<https://answers.yahoo.com/>

¹⁰<http://suggestions.yahoo.com/>

of question and answering systems are good examples of PSNs. They use crowdsourcing to gather knowledge and solve problems.

The other feature of Web 2.0 that facilitates crowdsourcing and emergent knowledge is the ability to add tags to any content. User generated taxonomy is utilized to categorize the pictures, videos, music and every kind of data. The tags make it easier to classify and categorize information and also to search and discover similar content on any topic (Wu et al., 2006). Collaborative tagging and bookmarking is used to add metadata and categories to pictures in Flickr and links in Delicious and it makes it easier for people to search and discover information. Using multiple tags also connects categories and geo-tagging helps in identifying any object based on location (Sinclair and Cardew-Hall, 2008; Xu et al., 2008).

Apart from collaborating together to create and share knowledge, people also build tools to solve problems. Instead of sharing information, motivated people come together and build projects and share it with the community. The Linux open source software project gave an alternative to the traditional software development and the programmers formed a community to build the software and tools. Linux provided the developers with a platform to collaborate and create where individual authorship is recognized, but they don't get exclusive intellectual rights. The Creative Commons license ¹¹ is used and it allows the creator to easily mark their creative work and share it with the world.

This type of large-scale software development requires lots of co-ordination and communication with the community and a project is divided into various phases of development. There are many tools available, like version control, bug tracking, task management and testing tools, for people to collaborate and communicate (Weber, 2004; Albors et al., 2008). Many software products like Firefox ¹², Android ¹³, etc. are successfully developed and used by general mass.

All of these examples suggest that the larger the online community, the more self-sustaining the system becomes.

2.2.2 Human Computation

Humans are good at solving the AI complete problems like natural language processing, image analysis that computers still find intractable. Luis von Ahn came up with various systems to utilize human computation and use crowdsourcing to solve these problems. The ESP games (von Ahn and Dabbish, 2004) are used to annotate images on the web (von Ahn, 2006), reCaptcha (von Ahn et al., 2003) is used to digitize millions of book and Duolingo (von Ahn, 2013) is used to translate the web into various languages.

¹¹<https://creativecommons.org/>

¹²<http://www-archive.mozilla.org/projects/firefox/>

¹³<http://source.android.com/>

These systems are used by users to play games, identify humans from machines to stop spam, and to learn foreign languages and they utilize the tasks easily done by human to solve computational problems. The same task done by many people gives consensus to knowledge and helps to prevent errors (von Ahn, 2009).

The Mechanical Turk provides a platform where people get paid to do human computation tasks. There are many research projects that ask users to perform annotation, alignment, sorting or editing tasks and people get paid for each completed task (Paolacci et al., 2010). There are also systems like CrowdSearch that adds a human layer to an algorithmic layer. This application combines automated image search with real time human validation of the search results (Yan et al., 2010).

The Foldit computer game takes advantage of human reasoning and puzzle solving abilities in prediction protein structure (Khatib et al., 2011). There are many citizen science programs like Galaxy Zoo that asks users to classify Galaxies (Lintott et al., 2008), Quantum Moves uses people to solve Quantum Physics problems, etc. (Silvertown, 2009).

All of these are example to indicate that crowdsourcing helps to solve problems. They use the wisdom of the crowd to find solutions. The success of the solution is dependent on its emergence from a large body of people solving the problem. This ‘wisdom of crowds’ is derived not from averaging solutions, but from aggregating them (Surowiecki, 2005). This crowdsourced information can be useful in searching for information by aggregation users’ responses and feedback. This helps to design PSNs.

2.2.3 Quality Management and User Recommendation

There are many user-rating systems that gather user’s votes and online behaviour. They use the aggregated data for quality management, monitor spamming and recommender system.

The power of collective behaviour and knowledge was utilized by websites such as Amazon¹⁴ and Pandora¹⁵ for recommendations of books and music respectively (Linden et al., 2003; Bu et al., 2010). Users of these websites also rate and review the things. The combined knowledge of the users’ rating and activity is used for predicting users’ needs. This is applied to other users and contents are recommended based on the aggregated user base data. YouTube and Netflix¹⁶ use viewers’ watching patterns to interlink the content and later recommend it to other users (Bennett and Lanning, 2007; Davidson et al., 2010).

¹⁴<http://www.amazon.com/>

¹⁵www.pandora.com/

¹⁶<http://www.netflix.com/>

On websites such as Reddit and Digg¹⁷ voting is used as a tool to popularise the content of the website and rank them higher so more people have access to it. Mechanical Turk and ESP games use the majority-selected answers to evaluate the correctness of the answers. Mechanical Turk also uses the user's work history to recommend task to help them find the right task (Ambati et al., 2011; Yuen et al., 2012).

The above examples show that user behaviour can be used to provide recommendations and do quality control in a system. Collective intelligence and wisdom of the crowd can be used to solve real world problems. The Web has enabled crowdsourcing and made it easier for people to collaborate and communicate. These ratings and recommendation given by users are useful in ranking information and search for information. This property of crowdsource system can be used in search and information retrieval as well as recommendations.

2.3 Semantic Web and Linked Data

This section discusses the concept of the Semantic Web. It describes what is the Semantic Web and what are different technologies used in Semantic Web. The main research question (RQ1) of this thesis is about using the Semantic Web and Linked Data technologies to solve problems in PSNs. It is vital to understand how Semantic Web technologies work and what tools are available to use. Another important aspect discussed in this chapter is how Semantic Web technologies help to solve problems on the Web, what are its benefits and how it can improve the current Web.

These technologies are used in the Suman system. This section is here to justify the use of technologies by providing example of systems and tools that use Semantic Web and how they use it. This knowledge is vital to design the Suman system. Furthermore, semantic search and query techniques will also be used to search for answers to unanswered questions in the PSNs, so understanding different semantic search techniques helps to design the Suman search algorithm.

2.3.1 Semantic Web Technology

The idea of the Semantic Web was coined by Sir Tim Berners-Lee (Berners-Lee et al., 2001). It described it as an extension of the current web, in which all data is structured and has a meaning accessible by both machines and people. Linked Data is a set of best practices for publishing reusable structured contents using the existent Web as a sustaining framework. In the Linked Data every resource is represented by a URI and HTTP URIs are used in order to retrieve documents that describe those resources. RDF is used to describe documents and they are linked to other resources (Bizer et al., 2009).

¹⁷<http://www.digg.com/>

The Semantic Web uses technologies of the current Web and formal semantics to overcome its limitations. In Web 2.0, webpages are written in HTML and does not have much metadata incorporated in it. HTML provides the markup to render the webpages element in the browser, but it does not provide meaning to the elements. It does not provide information if certain numbers are date or item number or price. The Semantic Web finds a solution for this problem by adding metadata and semantics to the documents so they are machine readable as well as human readable.

Berners-Lee described the Semantic Web architecture known as the Semantic Web Layer Cake or Semantic Web Stack as seen in figure 2.2. This model takes advantage of current Web standards as well as new W3C standards such RDF, OWL, etc. to create the structured and linked Web of data (Berners-Lee, 2003).

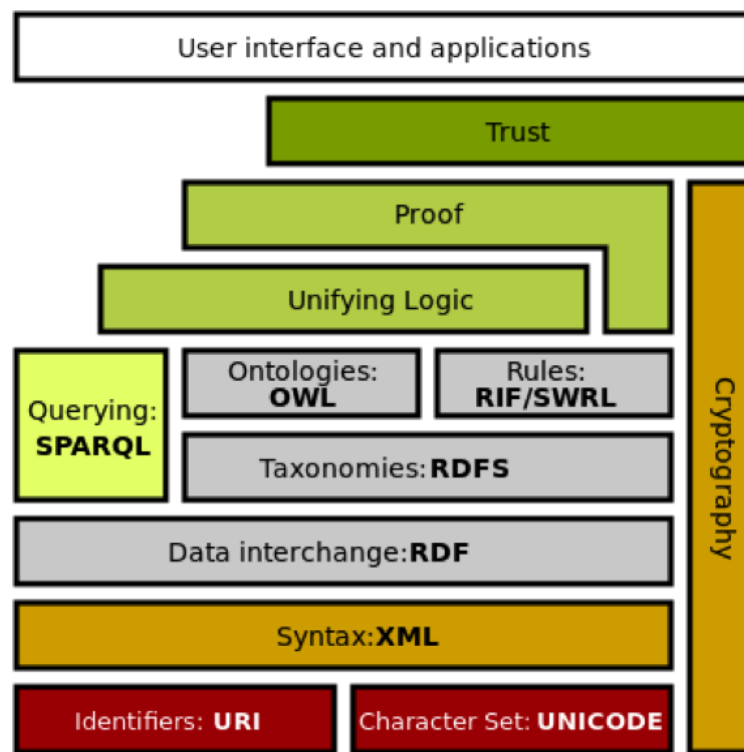


Figure 2.2: Semantic Web Layer Cake Model (Krauss, 2014).

As seen from figure 2.2, each layer exploits the capabilities of the layer below to extend the classical hypertext Web. The bottom two layers are examples of current technologies that form the base of the Semantic Web. The Semantic Web uses URIs to identify each resources and Unicode helps in encoding information in different languages (Berners-Lee et al., 2004). The XML layer uses XML namespaces and XML Schema to structure the content within a document. It also adds attributes and the schema of a particular domain in the XML document (Bray et al., 1998).

The middle layers use W3C standards developed to build the Semantic Web applications.

RDF is a framework for representing information about resources in subject-predicate-object triple format (Klyne and Carroll, 2006). It represents the data as a graph. It is used to store data as well as metadata about any resource. RDFS provides vocabulary to represent the schema in RDF (Brickley and Guha, 2000). It describes the properties and hierarchies of classes to create lightweight ontologies.

The next layer deals with logic and rules that take advantage of the meanings of the resources and infer meaningful information. OWL is an ontology language that adds more constructs over RDFS. It is based on description logic and allows addition of cardinality, properties and restriction of the values (McGuinness et al., 2004). RDFS and OWL can be used for reasoning within ontologies and knowledge bases. RIF is a rule interchange format that can be used to describe relations that cannot be described using OWL (Kifer and Boley, 2010). All the data can be queried using SPARQL query language (Prud'Hommeaux et al., 2008). SPARQL can query RDF data as well as ontologies and knowledge base described in RDFS and OWL. This is used to retrieve information for Semantic Web applications.

The top layer contains technologies that are not standardized by the W3C and are still work in progress as of 2015 (Hawke et al., 2015). The logic and proof layers will use all the semantics and rules from the lower layers and validate the deductions and inferences made. Formal proof with trusted inputs measured using provenance or other models will provide trust to the derived information. For reliable results cryptography methods such as digital signatures could be used to verify and ensure reliability of information. The top layer of user interface which will enable people to use Semantic Web applications.

Some of these technologies such as RDF, SPARQL, etc., are described in detail in the next section. They are extensively used in Linked Data and are more relevant to this thesis as they are used to build the Suman system.

2.3.2 Linked Data Technology

Linked Data is the publishing paradigm in which not only documents but also data items are linked to the Web. The linked datasets use Semantic Web technologies to structure the data and create relationships. The data are linked with different datasets and it can be queried to provide related information (Bizer et al., 2009).

Linked Data is an intricate part of the Semantic Web. Large-scale datasets are linked with each other and data publishers provide endpoints for query and reasoning. Berners-Lee has described four basic principles of Open Linked Data and Semantic Web. They are as follows (Berners-Lee, 2011):

1. Use URIs as names for things

2. Use HTTP URIs so that people can look up those names.
3. When someone looks up a URI, provide useful information, using the standards (RDF*, SPARQL)
4. Include links to other URIs so that they can discover more things.

These principles have been adopted in this thesis as a framework to create Linked Data. The interlinking of datasets and using HTTP protocol for discovery applies the general architecture of the Web to share data on a global scale. Hyperlinks enable users to navigate between different datasets and servers.

2.3.2.1 Benefits of Linked Data

Linked Data provides a flexible and generic data-publishing model. It facilitates publishing and consuming of the data. Data integration also helps to discover related datasets. There are many benefits of using Linked Data in the area of PSN and understanding these benefits justify the use of these technologies in the Suman system. They are (Heath and Bizer, 2011).

- **A unifying data model:** Linked Data provides a decentralized platform to link data across different domains. Data is represented by URIs and structured using RDF. The entities are described using globally unique identifiers. Hence, it can use different schemas to represent data. Multiple schemas can be used in parallel across domain and languages. RDF is useful to make the data machine-readable and read it across platforms. The data on the Web are usually represented in different format such as CSV, JSON, XML, etc. This requires different methods to integrate the data. RDF solves the data heterogeneity problem to some extent, semantic heterogeneity is still present.
- **A standardized data access mechanism:** A standard HTTP is used to access the Linked Data. Linked Data makes it necessary for the URIs to be dereferenceable. This allows the datasets and URIs to be easily accessible using a browser and APIs. This facilitates indexing and crawling of datasets by the search engines.
- **Hyperlink-based data discovery:** Linked Data offers a mechanism of linking datasets across different sources. This allows hyperlinks to be set between data that describe same entity or has some relationship with the data. When the dataset is queried, these links enable the application to discover new data sources at run time. All linked datasets create an Open Linked Data Graph bridging different sources.

- **Self-descriptive data:** Linked Data is structured using RDF schema and shared vocabulary. This metadata allows the data definition clear and vocabulary retrievable. Different vocabularies are also interlinked, hence providing more meaning to the entities.

PSNs can be benefited from using the Linked data because Linked Data provides an open platform, and a distributed and homogenous model to publish data. It makes it easier for data publishers and consumers to open, integrate and access data. It also provides useful endpoints for people to discover, reason and query the datasets to find related information.

2.3.2.2 Linking the Datasets

Linked Data follows the same principle as the Web. The Web links documents using hyperlinks into a global information space, similarly Linked Data link the different datasets into a global data space. Linking the datasets require following technologies (Heath and Bizer, 2011):

1. **URI:** URIs are at the base of Linked Data. URIs are used to identify web documents, digital contents, real world objects as well as abstract concepts. The abstract objects are like relationships between objects such as knowing someone (<http://xmlns.com/foaf/0.1/knows>) or other object's property such as colour.
2. **HTTP:** The second principle of Linked Data advocates using HTTP URIs. This makes the URIs dereferenceable over the HTTP protocol and provides more information about the objects. These descriptions can be in human readable HTML or machine-readable RDF. This is possible by using an HTTP mechanism called content negotiation (Heath and Bizer, 2011).

Content negotiation is a procedure according to which the HTTP clients send requests about the kind of document they prefer (HTML, RDF) in the HTTP headers with each request. Servers inspect these headers and provide an appropriate response of the document type.

There are two ways to make the real world object URIs dereferenceable. They are 303 URIs and Hash URIs. When a client makes a request to dereference a real world object URI, 303 URIs uses HTTP response code '303 See Other' response and redirects the client to a web document that describes the object. The client dereferences the new URI, and views the web document that describes the real world object. This is called '303 Redirect' (Sauermann and Cyganiak, 2008).

303 URIs require two HTTP requests to retrieve a single description of the real world object. Hash URIs solve this problem. The Hash URI contains a special part that is separated from the main URI by a hash symbol (#). This special part

is the fragment identifier. When client requests a Hash URI, the HTTP protocol requires the fragment part to be stripped off before requesting the URI from the server. The URIs that include hash cannot be retrieved directly. Thus, the main URI identifies the documents the hash URIs identifies the real world concepts without any ambiguity (Sauermann and Cyganiak, 2008).



Figure 2.3: The hash URI and 303 URI approach with content negotiation (Sauermann and Cyganiak, 2008).

The hash URIs have the advantage of reducing the number of necessary HTTP round-trips, which in turn reduces access latency. However, this approach has a downside. A client interested only in `#alice` will inadvertently load the data for all other resources as well, because they are in the same file. 303 URIs, on the other hand, are very flexible because the redirection target can be configured separately for each resource (Sauermann and Cyganiak, 2008).

3. **RDF:** Linked Data on the Web is represented using the Resource Description Framework. RDF is a framework for representing information in the Web in sets of subject-predicate-object triples. A set of triples is called RDF graph. The URIs representing the subjects and objects are the nodes in the graph and the predicates are the bridges that connect the nodes.

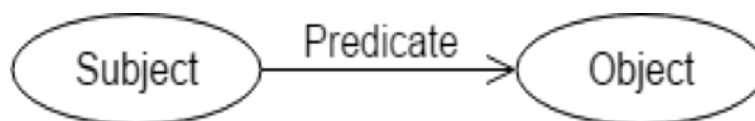


Figure 2.4: An RDF Triple.

The subject of the triple is a resource represented in a URI. The object can be a literal value such as number, strings, date, etc. or another URI describes another resource. The predicate describes the relationship between subject and object, e.g. name, location, date of birth, etc. The predicate is also identified by a URI that comes from ontologies and vocabularies. These vocabularies are generally standards and represent information about certain domain. RDF Schemas (RDFS) can be used to create more vocabularies about a domain that suits the data publishers to define their concepts and relationships. For example, FOAF vocabulary

can be used to describe user profile details, but there is no FOAF term to describe users' badges received on a website. This detail can be described by creating a new property using RDFS.

There are many benefits of using the RDF data model in the Linked Data. It uses HTTP URIs for resources as well as vocabularies, hence making it easier to scale. Clients can look up any URI in RDF graph to retrieve additional information. It enables linking between different datasets. Information from different sources can be merged together to create a global graph. RDF allows data to be represented using different vocabularies. This combined with RDFS and OWL gives freedom to structure the data as anyone desires.

4. **Linking:** The fourth Linked Data principle is to link the RDF data to other data sources on the Web. This external linking of datasets is fundamental to the Web of Data. It enables applications to discover additional information and link the data islands to create an interconnected global data space.

When URIs are dereferenced, the page shows an additional description of the data that might contain additional RDF links that can be dereferenced. A Linked Data browser can navigate this linked graph to find additional information.

There are 3 types of RDF links (Heath and Bizer, 2011):

- **Relational links:** They point to related things about the resources. This enables search of related background information about a resource. E.g. relationship links provide information about a person's date of birth, where they are from, etc.
- **Identity links:** This type of links enables aliases of resources to be linked together from different datasets. Identity links allow clients to find descriptions of resources from different sources and give different point of views. One example of this is when Owl: sameAs is used to link the aliases of any resources.
- **Vocabulary links:** This type of link connects the vocabulary terms to link to its definition and other related definitions. Vocabulary links make data self-descriptive and represent the data and the relationships. It helps to integrate data across vocabularies. owl:equivalentClass is an example of such linking.

Following and implementing all the four principles of Linked Data gives a 5 star rating to the dataset based on Berners-Lee five-star rating schema (Heath and Bizer, 2011).

In interest of addressing this research question these standards and definitions have been delineated for the purposes of implementation of these technologies in this thesis. The Suman system uses RDF to structure the data and uses Linked Data linking techniques to link the PSN data to the Wikipedia and OpenCalais datasets.

2.3.2.3 Linked Data Cloud

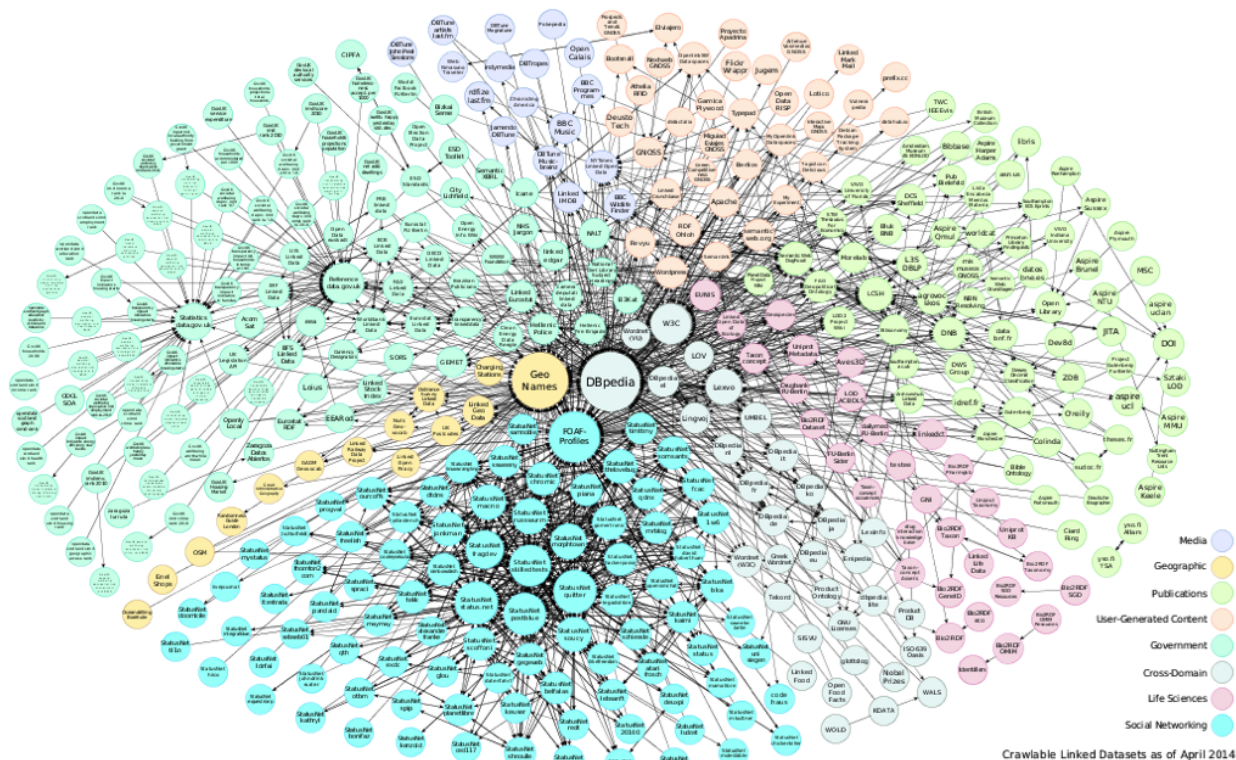


Figure 2.5: Linked Data Cloud (Schmachtenberg et al., 2014).

The Linked Open Data community project has the list ¹⁸ of all the open Linked Data which is connected together. The Linked Open Data Cloud as of April 2014 contains more than 1000 datasets. Figure 2.5 shows different interlinked datasets. The DBpedia dataset, which is the Linked Data format of the Wikipedia articles, is one of the biggest and most connected datasets (Schmachtenberg et al., 2014).

Berners-Lee has also coined the term Giant Global Graph to describe the Linked Data RDF graph (Berners-Lee, 2007). He considers it as important as the social graph. This is a decentralized graph in which everyone is free to link their datasets to another and finding related information. His Solid project will work towards the same goal. The Solid system is for building decentralized social applications using Linked Data (Conner-Simons, 2015).

The Suman system developed in this thesis links the PSN datasets to the DBpedia dataset and uses DBpedia categories to categorise the information. Understand the idea behind the Open Linked Data Cloud and how it works is vital in this thesis.

¹⁸<http://www.w3.org/wiki/TaskForces/CommunityProjects/LinkingOpenData/DataSets>

2.3.3 Semantic Web Vocabularies

There are many Semantic Web technologies available in the area of social network and social media. These technologies help with data representation, data portability and cross platform interoperability. Using Linked Data principles a decentralized system can be created that helps in social network integration. Every object and entity are identified using a URIs, and it is dereferenceable by using HTTP (Shadbolt et al., 2006). Also, useful metadata about the entities can be added in a structured format using RDF and could be linked with other related data to improve discovery of information.

As previously discussed there is already a large amount of Linked Data available on the Web called the Open Linked Data Cloud (Schmachtenberg et al., 2014). There are many useful ontologies and vocabularies to represent this data, including social data. Vocabularies define the concepts and relationships to describe resources. Vocabularies classify the relationship terms in a particular domain. In Linked Data vocabularies and ontologies are similar, there are no clear division between them. Ontologies are for more complex and formal collection of terms, whereas vocabularies is not used in strict formalisation. In Semantic Web and Linked Data vocabularies and ontologies are the basic building blocks for inference (W3C, 2015).

These vocabularies are used in this thesis to add structure and meaning to the data. They are used in the Suman system so understand the what these vocabulary describe and how it works is an important aspect of this thesis. Vocabularies help in data integration and remove term disambiguation. They also help in organizing knowledge and concepts in a domain. The evolution and formation of different and widely used vocabularies to describe people, communities, and their data are discussed below.

2.3.3.1 FOAF

The FOAF ontology is used to describe people and the relationship between them. Each person has a unique identifier and a specific vocabulary is used to create a personal profile of the users and describe their social network. FOAF describes people's personal details such as name, date of birth, email address, etc. and how they are linked with other people (friendship). It is easily integrated with other Semantic Web ontologies and a person can describe one or more of their online social network (Brickley and Miller, 2010).

Many websites such as LiveJournal, FriendFeed, identi.ca¹⁹ uses FOAF to describe their users. Many other websites have a plug-in or external application that creates their FOAF profile such as FOAF generator, WordPress plug-in, etc. FOAF helps to solve distributed identity problem when one user has different account in different website.

¹⁹<https://identi.ca/>

Using FOAF, a user can combine all the information from various sources into one file and interlink his different identity and social network. This also helps in creating a complete user profile and cross-site content recommendation (Bojars et al., 2008a).

Privacy and digital identity protection issues can also be resolved in FOAF by using SSL protocol that provides distributed authentication system to different user and create different policy for private and public information (Bojars et al., 2008b). FOAF + SSL authenticates a user in a connection made whenever accessing a web site. This is done by using the SSL layer built into a standard Web browser that implements HTTPS. In this approach trust is established recursively.

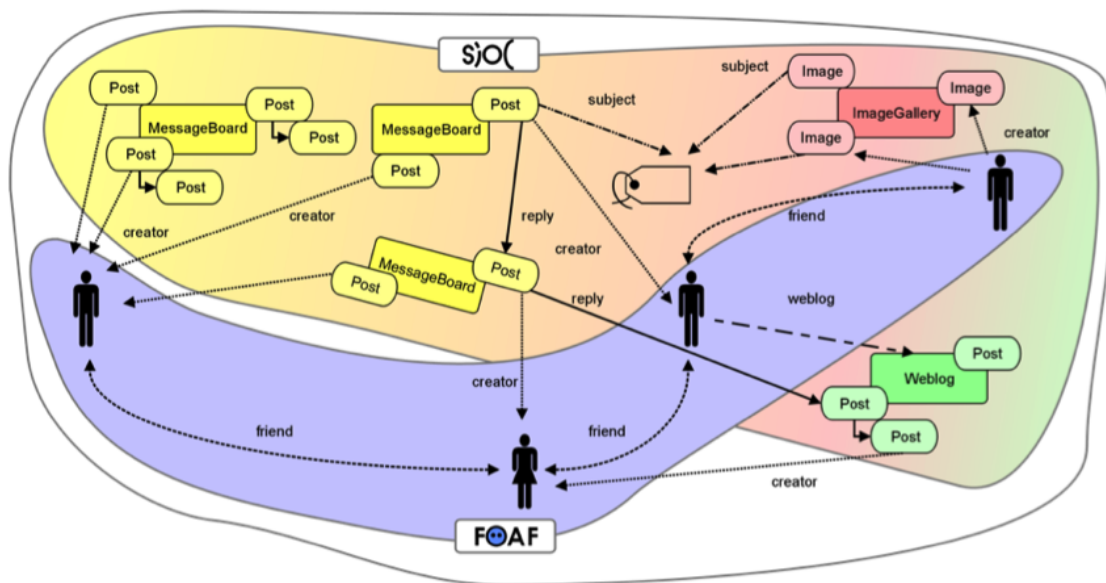


Figure 2.6: Creating a FOAF profile and SIOC file of a user.

2.3.3.2 SIOC

The SIOC ontology is used to describe online communities' activities and interlink the community data. Together with FOAF the combined document contains users' information, their social network and all the data user created with the comments and metadata provided by others (Breslin et al., 2005).

SIOC can be used for blogs, forums, discussion board, mailing lists, etc. Many websites use SIOC as well as many enterprises and E-government uses SIOC to freely available their data ²⁰.

SIOC is the next step in integrating different social networks together. It describes the content of the website in a structured format so the information is easier to access, search, and discover. The posts can be browsed by using specialized queries. SIOC

²⁰<http://www.w3.org/Submission/sioc-applications/>

links people from different networks. Together used with FOAF it can prevent identity problems or co-referencing. It allows people to connect the data in decentralized system and enable data integration (Breslin and Decker, 2007).

Figure 2.6 describes how FOAF and SIOC together can help interlinking and integrating different communities and different account of the same person together using Semantic Web technology for easy reusability. Both of them are implemented in this thesis, explained in detail in Chapter 4.

2.3.4 Social Semantic Applications

The Semantic Web technologies are used to add semantics to Web 2.0 social applications. These added semantics improve the linking and searching of information. The data has meaning and context, this helps to find the right information. The social semantic applications also showcase how adding Semantic Web technologies in Web 2.0 applications benefits the current web. PSNs can be benefited by studying these social semantic applications. These techniques can be added to current PSNs to add meaning, structure and context to the data. PSNs can itself become social semantic application if Semantic Web technologies are applied to it.

The Suman system aims to be a Semantic Web application and creates PSNs. Understand these applications gives insight on building and designing the Suman system. Some of the popular applications of Semantic Web technologies in the social web are discussed below.

2.3.4.1 Semantic Tagging

People use tags to classify, categorize and group their content using their own vocabulary. The websites get more user generate data; but this is also ambiguous and imprecise. People use different terms in the same context, meaning the same concept and same terms in different contexts for different meanings.

For example, if a post tagged with “Apple” is describing a fruit or a computer brand. It is necessary to reconcile lexical tags to URIs that unambiguously identify the meaning of a term (Garcia et al., 2009). The same tag actually differs when people use different spellings. There is also a lack of organization and hierarchy between the tags (Correndo and Alani, 2007). It is hard to be certain if a tag marked “Apple” is a subclass of a tag marked “Fruit”.

Most of these problems can be solved by the MOAT ontology that helps to describe the meaning of a tag semantically and connect it with the object and related tags (Passant

and Laublet, 2008). The co-occurrence of tags can be used to group similar tags together. It uses the collaborative approach to share the meaning of the tag in a community.

The Tag Ontology helps to solve the ontological and hierarchical problem that exists between groupings of tags. The SCOT (Social Semantic Cloud of Tags) ontology creates a model to describe a tag cloud and it makes the tag cloud portable so it can be exported from one service to another without losing the data (Kim et al., 2007).

Much research is carried out in the area of extracting ontologies from tags. FolksOntology is one of those projects that extracts relationships between tags (Van Damme et al., 2007). FLOR is another project that automatically identifies the meaning of a tag (Angeletou et al., 2008). These ontologies can be combined with other ontologies to create a complete model for semantic tagging in the area of social networking (Bojars et al., 2008a).

2.3.4.2 Semantic blogging and microblogging

Most blogosphere data is not structured or categorized, it is in plain text (Chin and Chignell, 2006). With the advent of Twitter and microblogging, this also lacks context and structure.

Semantic web technologies can be used to semantically annotate content and keywords of the blog posts and micro blogs to add meaning and structure to it. OpenCalais provides an easy to use platform to add annotation to the blog posts and it embeds RDFa into them (Corlosquet et al., 2009). The Twitter Annotation service takes the tweet metadata of location, time, etc. and annotates them to provide semantic metadata.

Another important project that supports semantic blogging is the Open Graph Project and Facebook Open Graph. These services help users to express preferences that improve the quality of the information provided in the blogs (Rowe et al., 2009). People can add reviews and ratings to a movie or use Facebook's "Like" button to show their preference and Open Graph connects the objects and resources with their friends and people with similar interest. It helps to provide better recommendations. These projects follow an Open Graph Protocol and create the data in RDF and the data can be queried to provide useful information.

The main use of semantic blogging technologies is that it is easier for users to publish their data in a structured format and they can publish it cross-site because it is portable. The developers can quickly create mash-ups to provide better data visualizations by integrating data from different sites because search and discover is easy in structured data. This data can also be mapped to people's FOAF profile and SIOC data files. If the dataset has a SPARQL endpoint, it can be queried to find more information (Millard et al., 2010).

2.3.4.3 Semantic Wiki

Wikis are a good example of collaborative creation and edition of emergent knowledge. There are a large number of people editing Wiki pages and helping in maintaining the quality of the information and preventing spam and irrelevant information (Chi, 2009). The structured information from Wikipedia info-boxes have already been converted into linked data within the DBpedia project, and this dataset is already one of the largest and most linked dataset in the Open Linked Data cloud.

DBpedia is very useful because it uses Wikipedia data and it is easy for anyone to read or write a Wikipedia article, and it also solves the problem machines cannot by using crowdsourcing. But DBpedia only uses part of the data because Wikipedia lacks proper structure and agreed semantics of the data and its categories (Auer et al., 2007).

The above problem can be solved by semantic wiki where all the data is structured and properly categorized. This provides better search and discovery of information like in OntoWiki. OntoWiki is a tool for agile, distributed knowledge engineering. It uses a similar approach as the Wiki for browsing and authoring of RDF knowledge bases. It offers an information map views of the stored information and an online editing mode for RDF data. Social collaboration is facilitated by keeping track of changes, allowing to comment. People can also rate and measure the popularity of content. Ontowiki enhances the browsing and retrieval by offering semantic enhanced search strategies (Auer et al., 2006).

Wikidata is another tool used within the other Wikimedia projects to provide well-maintained, high-quality data. Wikidata acts as a secondary database to Wikipedia. This means that instead of containing facts, it contains references for facts. This includes inconsistent and contradictory facts. This helps in representing the diversity of knowledge about a given entity. It supports multilingual datasets (Vrandečić and Krötzsch, 2014).

2.3.5 Semantic Search and Query

One of the strengths of the Semantic Web and Linked Data is in searching the related information based on different categories and concepts. Semantic search uses the contextual meaning and relationships of the keywords for information retrieval. It evaluates and understands the meaning of key phrases to find the search results (Guha et al., 2003).

Semantic search is the combination of the conventional Information Retrieval (IR), web search and knowledge management methodologies. Semantic queries enables the retrieval of derived information based on semantic and structural information contained in the data. The Suman system uses semantic search techniques to search for answers

for unanswered questions. Hence, understanding different type of search techniques is an important part of the background literature review.

2.3.5.1 Information Retrieval

The traditional Information Retrieval (IR) methodologies are based on the occurrence of words in the documents. This has been used in libraries, organizations and many document systems. This approach mainly uses keywords based queries and returns a list of relevant documents with those keywords with different degree of relevancy.

Some of the classic and widely used IR models are the Vector-Space model (Salton et al., 1975), Probabilistic model (Eddy et al., 2009), Latent Semantic Indexing (Deerwester et al., 1990), Machine Learning based models (Sebastiani, 2002), etc. Many search systems uses some form or combination of these models (Kosala and Blockeel, 2000).

Term frequency-inverse document frequency (tf-idf) is a widely used syntactic measure to determine the importance of a word based on the number of occurrences in a document (Salton et al., 1975). This is relative to the number of occurrences in the entire collection of the documents. The Vector Space model represents text documents as vectors and relevance ranking of documents in a keyword is calculated.

A probabilistic model calculates the probabilities of similarities of the documents to the query (Eddy, 1998). The Bayes theorem is often used as probabilistic inference to retrieve documents. Latent Semantic Indexing uses singular value decomposition techniques to identify relationships between keywords and concepts. Latent Semantic Analysis finds the concept of the words and correlate semantically related terms (Deerwester et al., 1990). This method retrieves conceptually similar documents to the search query.

Machine learning algorithms have been applied in the area of IR. (Langley and Simon, 1995) had identified five major paradigms that can be applied in this area. They are rule induction, instance-based learning, neural networks, genetic algorithms and analytical learning. These algorithms apply heuristics to generate structures that represent the relationships implicit in the data.

These are some of the important types of IR models and gives a general overview all different approach to solve similar problems. There are many more models present, but in this thesis and the design and development of the Suman system a similar approach to tf-idf model has been taken so it has been discussed here.

2.3.5.2 Web Search

These IR technologies, discussed above, have also been used in the Web. The experience of the Web improved significantly when the search engines started to crawl the Web

and provided a centralised portal to discover any webpages and information. Search engines augmented these existing methodologies with the hyperlinks and started to rank the search results using the PageRank (Lawrence Page and Winograd, 1999), HITS (Chakrabarti et al., 1999), Citation Indexing (Lawrence et al., 1999), etc. algorithms.

Conventional search techniques are developed by augmenting word computational model and enhanced by the link analysis (Jiang and Conrath, 1997). Some of the current popular models used in the web page retrieval include the combination of content-based approach and link analysis methods. The content-based approach uses the IR methods to analyse the content of the web pages to find the best matches to the search query (Eiron and McCurley, 2003). The link analysis method is similar to the citation analysis used in the area of bibliometrics (Garfield et al., 1972). It analyses the hyperlinks that link the web pages, both in links and out links. It uses the Web's graph structure to determine the importance of the webpages and ranks the results accordingly. There are ways to detect the link spams to filter out the lower quality content (Bharat and Henzinger, 1998). (Li et al., 2012) also describes the majority of links on the webpages are to organize the information, not for the recommendation. A path-based model is used to distinguish the information organization and recommendation to rank the search query. There is a lot of research available in this area that describes the use of Web's graph structure to improve the web search query (Lawrence Page and Winograd, 1999; Kleinberg, 1999).

There are certain limitations to the current web search methodologies. The text-based search provides results for exact keywords and phrases. This sometimes does not provide the search result for users who don't know what exactly they are looking for. This approach provides limited results in the research based queries as described by (Guha et al., 2003). The availability of well structured machine understandable information and datasets offers opportunities to improve the traditional search methods (Guha et al., 2003) and it could be Incorporated in PSNs.

The Suman system tries to overcome some of the issues and limitation of the web search approach and these literatures gives general overview of different types of algorithms and techniques used in web search. Some of these techniques are applied in the Suman system as indexing and ranking mechanism. This section gives a summary of important aspects of web search and what different approach could be used in the Suman system.

2.3.5.3 Semantic Search

Semantic based IR focus on understanding the meaning of the document instead of calculating the frequency of any words appeared on the documents. The Semantic search focuses on understanding the concepts of the objects and finds meanings in the structured

data (Guha et al., 2003). It utilises domain knowledge and ontology navigation to modify the query and apply context of the topic to search (Castells et al., 2007).

There are different research areas in semantic search. There have been some attempts to classify and categorise them (Mäkelä, 2005; Mangold, 2007). Understanding these different techniques helps to choose the right approach for the Suman system. This also helps to figure out what could be the right approach to use based on the dataset available and the outcome required for the Suman system. A brief summary of all the different research methodologies is below.

- **Keywords to concept mapping** - Early research in the Semantic Web added meanings and structure to text using an ontological approach or by finding similarities between words. (Li et al., 2003) and (Freitas et al., 2013) proposed methods to find similarities between keywords by using the WordNet (Fellbaum, 1998) thesaurus ontology navigation to expand the query. Other approaches also use Wikipedia categories (Han and Zhao, 2010). There are many systems that get synonyms and meronyms for keywords by graph traversal and utilize the concepts to broaden or constrain search term (Moldovan and Mihalcea, 2000; Buscaldi et al., 2005; Varelas et al., 2005; Liu et al., 2004). Some other example of these systems are TAP, Clever Search, etc. In TAP (Guha et al., 2003), keywords are matched against concept labels in an RDF repository. (Rocha et al., 2004) uses text search first, then the RDF graph traversal finds related concepts using spread activation algorithm. (Airio et al., 2004) and (Grootjen and Van Der Weide, 2006) uses the same approach, but user interface allows people to browse the ontology.
- **Ontological based search** - Data on the Semantic Web is divided into ontological data and instance data. The domain knowledge and relationship is described as class relationships and the actual data is an instance of the class. This kind of search deals with efficiently locating the instances of the class. In SHOE search system user construct queries by attribute value of ontology classes (Heflin and Hendler, 2000). SEAL uses a class subsumption tree based approach similar to Yahoo (Maedche et al., 2003). OntoViews-based portals (Mäkelä et al., 2004), and (Vandic et al., 2012) use multifaceted search approach. Here multiple distinct views are provided into the data. Other examples are (Tran et al., 2007), (Ding et al., 2004), (Kara et al., 2012).
- **Graph patterns**- Many kinds of complex queries can be formulated to find certain groups of objects connected by certain relationships. This is possible due to the graph patterns in the datasets in which the objects are node and relationships and the arcs that join them (Zeng et al., 2013). The system also uses ontologies as directed graphs and exploit links between entities to provide navigation. The main problem with this is the queries are not easy to formulate (Ferré and Hermann, 2011). SWSE is a graphical user interface for building graph pattern

queries that are based on navigating the underlying ontology (Hogan et al., 2011). Multifaceted search portals provide a user interface for creating a very constrained subset of complex graph patterns. CS-Aktive used Linked Data structure to provide a unified view of distributed dataset and used reference ontology as mediator (Shadbolt et al., 2004). RKB Explorer applied similar methodologies, but it used heuristic methods to get URI equivalence (Glaser et al., 2008). Hence, it shows consistent reference service.

- **Logic and inference** - One of the fundamental advantages of using ontological knowledge on the Semantic Web is making inferences based on rules to solve problems. Only simple applications have been created to apply this, actual implementation of this is quite rare. Rule-based systems are hard to implement because the Web works based on an open world assumption, but well explored logic works in a closed world. Pellet (Sirin et al., 2007) and OWL-QL (Fikes et al., 2004) are DL-reasoners and they are simpler versions of this. OWL-QL uses if-then queries to reason. (Koutsomitropoulos et al., 2011) uses entailment based queries and captures the meaning of the query in reasoner-compatible format. The IRIS system is for repositories to search for articles (Wei et al., 2007). It uses an inference engine for reasoning with rules of computer science domain knowledge where broader and narrower terms are defined in SKOS.
- **Emerging Knowledge**- Graph patterns and property relations can be used to traverse from one resource to another using the connected path. Analyzing the relationships between resources sometimes could create an emerging knowledge. This is a hard problem to solve because the resources need to have interesting links between them, but they need to be general enough to have complex hidden relationship in the data. The Flink system helps find prominent researchers and experts in an area (Mika, 2005). It uses extraction, aggregation and visualization of social communities and professionals to find the experts. Ontocopi uses breadth first and spreading activation search to identify communities (Alani et al., 2002). Another example is Arnetiminer that uses PLSA model to rank experts (Tang et al., 2007).

The Semantic Web is a wide area of research and there are many different systems that combine multiple methodologies to get the best results. Semantic Web technologies apply tools to find simple concepts in the data and create relationships. This graphical nature of the objects combined with fuzzy logic and keyword search results in creating complex queries. These ideas have been included in this thesis and it has been applied in the Suman system. All the approaches that are part of the Suman system design has been discussed in detail in the next sections.

2.3.5.3.1 Keyword Disambiguation: Semantic based IR methods tend to understand the meaning of the keywords to match the best documents to the queries. This is done by adding semantic tags into the texts of the documents and queries. This is a hard problem to solve because of the presence of too many non-relevant items that provide false positive results and exclusion of too many relevant items because of the effects of synonym and polysemy causing false negative results (Guha et al., 2003).

The Semantic Web solves the problem of keyword ambiguity by adding ontology classes or other semantic tags to define the keywords. The keywords can also be extended using fuzzy logic to add all the variations of spellings and synonyms and meronyms. WordNet and Wikipedia are good source to solve the keyword disambiguation (Moldovan and Mihalcea, 2000; Varelas et al., 2005). Many systems like TAP (Guha et al., 2003), AquaLog (Lopez et al., 2007) uses these services to extend their knowledge of keywords.

This thesis deals with the technology related documents where many terms are similar to common vocabulary in English. For example Java is a programming language as well as coffee beans from Indonesian island of Java. It's important to tag the keywords properly to find the right results for queries that ask about coffee and that ask about errors in their program written in Java.

There are many approaches to solve the keyword disambiguation problem. This field was built on the work of the computational linguistic community (Wilks and Stevenson, 1997; Schütze, 1998) and extended by the Semantic Web community (Moldovan and Mihalcea, 2000; Trillo et al., 2007; Varelas et al., 2005). SemTag uses Taxonomy-Based disambiguation algorithm to annotate large scale webpages (Dill et al., 2003). It applies TAP knowledge base to find the context of the keywords. The platform developed by (Trillo et al., 2007) implements a similar approach, but uses Swoogle (Ding et al., 2004) and WordNet knowledge base. It utilizes their ontology pool to extract sense of the word.

There are many machine learning approaches to solve keyword disambiguation problem. (Wang et al., 2012) used unsupervised learning methods. They relied on predefined similarity matrix and heuristic rules to find the match. But supervised learning method is widely adopted approach where the algorithm infers the function using a training dataset. (Zheng et al., 2012) and (Mendes et al., 2011) used designed supervised classifier to select and rank the best target. (Nebhi, 2013) and (Zheng et al., 2012) used the Freebase, the vast entity database to solve the entity disambiguation using semi supervised machine learning techniques.

There are systems like Wikipedia Miner (Milne and Witten, 2012), Dbpedia spotlight (Mendes et al., 2011) that uses Wikipedia and Dbpedia data to annotate text, and disambiguate the main topics in the text. They make use of the rich hyperlinks available in the articles to compute similarities. They also use spotting algorithm to find the context of the text and then map the spotted phrase with the Wikipedia article. Other

systems that use Wikipedia are (Bunescu and Pasca, 2006) and (Cucerzan, 2007) , they use the Wikipedia categories to improve the accuracy. (Corlosquet et al., 2009) use OpenCalais dataset to resolve name disambiguation for their content management system.

In the Suman system existing tools Wikipedia Miner and OpenCalais are used to perform keyword disambiguation.

2.3.5.3.2 Concept mapping Another widely used approach to attempt to improve the semantic search is linking the keywords to concepts and categories (Bhogal et al., 2007). This provides with broader and narrower terms to query and helps find better search results by exploiting the keywords and concept relationships (Trillo et al., 2007).

The Linked Data and Semantic Web community usually use the open datasets to map their keywords with the existing, well defined concept knowledge base. WordNet, Wikipedia, Freebase are the important dataset as discussed in the section above.

(Li et al., 2014) and (Buscaldi et al., 2005) used a semantic query expansion approach by using the concept relationships from Wordnet and Wikipedia. (Grootjen and Van Der Weide, 2006) and (Bhogal et al., 2007) used the domain ontology knowledge for query augmentation and substitution. The SHOE search system implements the ontology hierarchical knowledge to provide users with more filters for search (Heflin and Hendler, 2000). The SWED directory portal applies multi-faceted search approach to navigate through different concept hierarchy (Matthews, 2005).

The Vector Space model was adapted to improve the ranking of documents by exploiting the ontology based knowledge (Fernández et al., 2011). (Rinaldi, 2009; Hao et al., 2008; Egozi et al., 2011) computed the semantic relatedness between defined concepts to rank the documents. The graph formed by the linked concepts can be used to find the relevant document by measuring the relevant graph path and the importance score is measured by calculating the relevance between queries and documents (Brauer et al., 2009).

Co-occurrence of keywords can provide insight into the similarities of concepts and can be utilised to rank documents and search results (Rocha et al., 2004). (Lamberti et al., 2009) and (Kara et al., 2012) uses a document semantic annotation graph to find semantic relationship between concepts for query matching. RKB Explorer uses the SameAs property to find similar concepts and aggregate all the concepts to show consistence reference (Glaser et al., 2008).

The Google knowledge graph (Singhal, 2012) and Wolfram Alpha (Wolfram, 2009) uses CIA The World Factbook (CIA, 2010), Freebase and Wikipedia to their concepts. Bing has its own Satori knowledge base. These are widely used applications of the concept-mapping methodology.

Concept mapping is also implemented in this thesis to expand on search terms in PSNs. All the concepts in the Suman system are mapped to Wikipedia and Open-Calais datasets. Wikipedia categories are also used to categorize the concepts in the Suman system to expand the search terms.

2.3.5.3.3 Semantic Queries Semantic query is used in Linked Data and in some cases on the Semantic Web for IR (Ferré and Hermann, 2011). They enable retrieval of both explicitly and implicitly derived information. It takes advantage of semantic data structure and added semantics and relationships. The query engine also performs pattern matching and reasoning to find answers to wide open questions.

SPARQL is a popular query language to formulate a semantic query in similar syntax as SQL (Prud'Hommeaux et al., 2008). It works on named graphs, triples or linked data. Named graphs are graphs in which a set of RDF statements are identified using URIs. This allows descriptions to be made of that set of statements such as context, provenance information or other such metadata. It is possible to use the graph structure to process the relationships between the objects and infer answers. General semantic search works on unstructured data with semantics, but SPARQL query requires structured data represented in RDF, turtle, N-triple, etc. It is like a database query that requires precise relational-type operations. Hence it can perform features like operators (greater than, less than, equal to), pattern matching and namespaces. It can also apply semantic rules, transitive relations and ontological subclass and perform contextual full text search. SPARQL engines also use graph traversal techniques to derive new information.

RDF is stored in the database known as triplestore. Triplestores provide a flexible model to store graph data. Some of the triplestores are made by extending relational databases. They have an intermediate ETL process to Extract, Transform and Load the data (Rohloff et al., 2007). Some popular triplestores are Virtuoso ²¹, 4Store ²², Stardog ²³, etc. The data provider sometimes provides a SPARQL endpoint to query the database. There are also Jena ²⁴, Sesame ²⁵, Oracle RDBMS ²⁶ based data storage solutions that provide a complete framework to store, process and make inferences based queries. (Cyganiak, 2005) presents a relational model for SPARQL to uses relational algebra (join, projection) for selection. Federate is a relational database gateway that maps RDF queries to existing database structure (Prud'hommeaux, 2004). Jena is a JAVA toolkit that provides complete support to manipulate RDF models (McBride, 2002).

²¹<http://virtuoso.openlinksw.com/>

²²<http://4store.org/>

²³<http://stardog.com/>

²⁴<https://jena.apache.org/>

²⁵<http://www.sesamedatabase.com/>

²⁶<http://www.oracle.com/technetwork/database/options/spatialandgraph/overview/rdfsemantic-graph-1902016.html>

There is also Ontology Based Data Access (ODBA) systems that allow querying relational databases through conceptual representations mapped to an ontology of the domain (Kogalovsky, 2012). ODBA allows SPARQL queries, R2RML mappings, etc. D2RQ, Ontop, are some of the ODBA systems. Other ways to query the database is the use of graphical search. GRQL is a graphical user interface for building graph pattern queries that are based on navigating the ontology (Athanasios et al., 2004).

There has been a lot of research in the area of query optimisation and query construction to find the best results (Schmidt et al., 2010; Hartig and Heese, 2007; Tsialiamanis et al., 2012; Stocker et al., 2008). Different query engines apply different methods or combinations of methods to search and rank results. The DARQ engine decomposes the query into subqueries and forward them to the query service (Quilitz and Leser, 2008). It combines the results received from the subqueries to show the final results. SemWIK works similarly and has a mediator service to distribute the queries (Langegger et al., 2008). Tabulator, a Linked Data browser, traverses RDF links to obtain additional related information about the resources (Berners-Lee et al., 2006a).

The SPARK system uses term mapping, query graph construction and query ranking to adapt the keyword queries into a formal logic queries (Zhou et al., 2007). Ontolook is another keyword based search application that infers the possible relations between the keywords to improve the precision of the search (Li et al., 2007). SemSearch has a structured keyword query interface to hide the complexity of the system (Lei et al., 2006). It takes natural language queries and turn it into formal queries for reasoning. Keyword entity is then matched against subject, predicate or objects in the knowledge base. The Avatar Semantic search engine takes advantage of the annotations in the classical keywords based search methods (Kandogan et al., 2006). (Tran et al., 2007), (Wang et al., 2008), (Rocha et al., 2004) has also adapted keyword based query into their corresponding SPARQL based formal queries. The Quick system allows users to incrementally add keywords to search any information in a domain to construct a query (Zenz et al., 2009). It combines the convenience of keyword queries with the precision of semantic queries.

The next step after making queries is ranking the results. (Li et al., 2014) has added two steps document ranking to the initial keyword based search to improve the results. The limitations of keyword-based search are sometimes overcome by exploiting domain ontology of the knowledge base. (Fernández et al., 2011) used vector space model approach to find related ontology and improve document ranking. (Hao et al., 2008) and (Rinaldi, 2009) computes the document relevance by comparing the similarities of words using ontology. (Daoud et al., 2010) and (Brauer et al., 2009) converts the free text content into semantic graphs and use a graph matching algorithm to rank documents. (You et al., 2009) considered queries as concepts and documents as instances, then use ontology reasoning to calculate document relevance. These models uses semantic relations

defined by the ontology for query expansion or semantic similarity calculations and then rank the documents.

The keywords only search approach is popular because of ease of retrieving information. However, it lacks the in depth knowledge of user search intentions. It also does not have enough expressivity (concepts and topics) in the search query. This could potentially result in less effective ranking of results when user's intentions are not completely clear. The hybrid approach to keyword search minimizes this issue (Rocha et al., 2004). This is seen in different cases when a keyword search is extended into extended query terms using ontological knowledge, or using graph traversal to find related objects. Previous research suggests that semantic linkage when added to keywords based search improves the accuracy of the search results.

Understanding these different approaches gives insight on how different system works and how it can be implemented in the Suman system. The Suman system uses SPARQL query engine called the Stardog and it uses the Lucene search engine for indexing and querying. The details of the Stardog query engine and how it works and how the indexing is improved is discussed in details in Chapter 4.

2.3.5.3.4 Expert Recommendation: There are many recommender systems available and it is a huge area of research. There are many papers (Bobadilla et al., 2013; Felfernig et al., 2013; Park et al., 2012), that give a good insight on the different recommender system and models available. Earlier, the recommendation systems were based on demography, user behaviour and collective filtering. These recommender systems could be divided into three types (Bobadilla et al., 2013)- Content-based recommendations where the user will be recommended items similar to the ones the user preferred in the past; Collaborative recommendations where the user will be recommended items that people with similar tastes and preferences liked in the past; and Hybrid approaches where these methods combine collaborative and content-based methods. Nowadays, many other factors are incorporated like social information, location and other personal information gathered while user profiling. These system use many approaches from tf-idf, clustering or nearest neighbour.

In this thesis, the area of expert recommendation is studied that uses Semantic Web and Linked Data technologies. There are many recommendation systems that use Semantic Web techniques like Sem-Fit for tourism (García-Crespo et al., 2011), CHIP for digital museum collection (Aroyo et al., 2007). But this thesis focuses on expert recommender in PSNs. (Fazel-Zarandi et al., 2011) uses social network analysis and data representation techniques to map users to the domain data and recommend experts in the scientific field. (Venkataramani et al., 2013) ranks the experts by mining their activity on multiple domain like StackOverflow and Github to build user model and recommends experts based on their activity. (El-Korany, 2013) does something similar, they use a cascade

model to aggregate knowledge extracted from various online communities and use the Vector Space model to find relevant content and use PageRank to rank the experts in a community.

In TEM, probabilistic generative model is used to model topics and experts. They rank users based on interest and expertise on different topics and recommend them (Yang et al., 2013). (Davoodi et al., 2013) used Wikipedia knowledge to build a hybrid expert recommendation system that integrates the characteristics of content-based recommendation algorithms into a social network-based collaborative filtering system. (Stankovic et al., 2010) used Linked Open Data to link people from different platforms based on their activity and expertise and recommended them.

Similar approach is taken in the Suman system to generate user model based on users' activity on PSNs. Their interest is aggregated based on their post history and they are linked to Wikipedia and Linked Open Data.

The details of the Suman system, the semantic search and expert recommendation techniques used can be found in Chapter 4. The literature discussed above gives an insight of similar applications that uses similar approaches to the Suman system. It gives support to the design choices made and provide information about other research that are similar and can be useful.

Chapter 3

Purposive Social Network

A community is created when people collectively create and share information to provide a common source of knowledge (Lazar and Preece, 1998). The Web provides a virtual environment for people to create an online community. The communities have a purpose, are supported by technology, and are guided by norms and policies (Preece, 2000).

Online communities provide an easy platform for people to come together and participate. The other emerging trend on the Web and online communities is using the power of the crowd to create knowledge. As discussed in section 2.2 community projects like Wikipedia provides a platform for people to come together and create a shared knowledge base. The successful online communities depend on users' participation, efficient moderation, interest in the topic and common ties between communities members (Bishop, 2007).

The communities where people with common interests come together and solves problems, and create a self-sustaining and self-regulating community are defined as a Purposive Social Network (PSN). Defining PSN using the scientific literature and existing example of PSN will answer KQ1.

In this thesis question and answer forums for the programmer, StackOverflow and Reddit, are studied as examples of PSN. The analysis of these two websites in the context of PSN, studying it's community structure, user interaction and behaviour to find the motive and incentives used will answer KQ2.

In this chapter PSN is defined in more details, its characteristics and benefits are studied in the context of StackOverflow and Reddit.

3.1 What is Purposive Social Network

There are many reasons to join and form communities; human are social creatures and forming a social tie is a natural instinct (Cooley, 1992). In the online world, people from any part of the world can connect together and have a certain degree of anonymity and privacy (Bernstein et al., 2011). People can reach to others with similar background and interest.

A Purposive Social Network can be defined as a social network with a purpose, as the name suggests. It is created when people come together with a common interest and objective to share information, build a knowledge base, solve problems or achieve some common goal and purpose. The purpose could be defined by the community guidelines or people in the community themselves (Preece, 2000).

PSN is an social network because people form ties with each other. It could be strong ties or weak ties. The relationship between the people could be explicit when they themselves create friendships, or implicit where they interact with each other.

The main feature of PSN is the problem-solving component. People form community to solve a common problem. They have a clear goal and work towards the same objective. People in the community have the similar purpose and they work towards solving a common problem. PSN varies from other social networks in this regard. People communicate and work together towards the same goal, not just to communicate ideas with each other (Brabham, 2008; von Ahn, 2009).

PSN is a homophily. People in the community have similar interests. They work at the common problem and share common interest. This may be a heterogeneous network involving individuals from different locations and background. But they have a shared focus and shared interests (Monge and Contractor, 2003; Garton et al., 1997; McPherson et al., 2001).

PSN could be a communication network as well as a knowledge network. A communication network is analysed by studying the communication links used to seek information from knowledgeable others. Knowledge networks are the mechanism through which people share explicit and tacit knowledge (Monge and Contractor, 2003).

The emergence of communication network can be linked to and embedded in a knowledge network. Groups of people share knowledge and information with each other. This might lead to an emergent knowledge base (Monge and Contractor, 2003; Kogut, 2000; Zettsu and Kiyoki, 2006).

PSN relies on users to solve the problem and share knowledge. It could use the wisdom of the crowd for its purpose (Kittur et al., 2007; Shang et al., 2011). There are many systems that can be PSNs. This thesis focuses on the question and answering systems

– StackOverflow and Reddit. In these websites people answer each other's queries and solve programming related the problems using crowdsourcing.

3.2 Different types of communities in Purposive Social Network

PSN is a term defined to describe certain types of communities. Most of the research in the community is done on detecting the structure or the communities (Newman, 2004; Flake et al., 2002; Chin and Chignell, 2006). Some researches that categorize the community based on user behavior, privacy and community structure (Zhang et al., 2007; Plant, 2004; Angeletou et al., 2011).

This section categorizes communities based on their role and how the community provides an objective to users to solve problems (Lazar and Preece, 1998; Backstrom et al., 2006). These communities can have many features in common, but they might have some distinct features of their own that might not be mentioned in this section. Understanding different types of communities that could be PSNs help to answer KQ1 better. It should be noted that a community can be of one or many categories mentioned below.

3.2.1 Information Based Community

An Information Based community is a community that shares any particular type of information or resources. In these communities people come to share their knowledge and gather more information on any matter. IRC channels, forums, mailing list are some examples of such community. The communication is fast (IRC) or delayed (mailing list) based on the communication medium.

The main feature of this community is information diffusion. Contents are the focus of the community. Nowadays, with the popularity of Twitter and Tumblr, it is easier to share information using retweet and reblogs. Here, the network cascade effect is strong which means that the information can pass through one person to another in their social network. This flow of information facilitates the information going viral (Kwak et al., 2010). This kind of community generally has many new people joining to get some information but the rate of return is low (Tantipathananandh et al., 2007).

3.2.2 Interest Based Community

An Interest Based community is a community where users come together to share their hobbies or interests and form a network. The number of people in the network is generally stable and they actively participate in any discussion. Online role-playing

and gaming community are a typical example of this type of community. Researchers are taking advantage of skills and interest of the gaming community to solve scientific problems through human computation (Khatib et al., 2011; von Ahn and Dabbish, 2004). Foldit and Planet Hunters are some of the projects that take advantage of these communities.

Forums and boards are another good example. They provided a place for people to meet anonymously and discuss their issues. This was the beginning of virtual community where face to face communication was replaced by messaging boards, chat rooms, question and answer forums. These online communities were formed based on interest and common objectives and people from all around the world could come together and communicate about a particular topic of interest (Adamic et al., 2008).

3.2.3 Expert Based Community

An Expert Based community is formed by experts in a field. Experts come together in order to discuss a common matter or to solve a common problem. They do not really have to have complete knowledge of the topic, but a good understanding of the subject. They might have previous experience with the topic that they share with the community. The communities can also be considered as knowledge networks where the knowledge flows from one person to the next. This type of community puts a lot of emphasis on the quality of the knowledge shared among the community (Zhang et al., 2007; Zettsu and Kiyoki, 2006).

StackOverflow is a good example where experienced developers answers programming questions. They have different sites for different expert communities. Wikipedia is also an emergent knowledge base where people with expertise on a topic contribute to it (Kittur et al., 2007).

3.2.4 Location Based Community

A Location Based Community is a community where people come together for a geographical reason. This type of community is created where proximity is important. These types of communities form because of social issues concerning a region.

Neighbourhood watch, Yoga and running groups or support groups are examples of this type of community (Smith, 2008). They are region based and people not only interact online, but also physically.

StackOverflow and Reddit are examples of expert based communities are data from these websites are used in the Suman system to study them in detail.

3.3 Properties of Purposive Social Network

There are no fixed types and features of PSNs. But they have certain features that are prominent. The network structure of the PSN communities are same as any other social network, but some attributes and properties make them different (Katz et al., 2004; Plant, 2004). These properties they can have but it is not necessary, they will always have them. It is not intrinsic properties, they are subjective properties that a PSN may or may not have them. Understanding these special properties of PSN helps to answer KQ1. They are as follows:

3.3.1 Community Size

People with similar objectives and purpose create a PSN community. It does not have to consist of a large number of people. The community size varies across different websites and their user base. It is often that small number of people come together to share knowledge and solve a problem.

For example, popular tags in StackOverflow and popular subreddits have hundreds of thousands of users in the community, but a relatively niche community might only have a few hundreds users associated with it.

3.3.2 Focused Interest

A PSN community focuses on a common objective. Hence, sometimes the community does not have a broad range of interests or activities associated with them. They are not aimed at everyone. Small groups of people with shared interest come together to solve a very specific problem. Usually, they have focused interest and they create a focused knowledge. E.g. Reddit has specific subreddits for different topics and each subreddit has its own community of users.

3.3.3 Direct Communication

A PSN is generally a communication network. People do not have strong and explicit ties with each other (Monge and Contractor, 2003). However, it is easy to communicate with people in the community. Each individual member in the community shares information and communicate directly with other members. They either broadcast the message for everyone to see or communicate directly with one another.

3.3.4 Active Participation

A PSN communities have focused goals and interest and number of people in their network vary. In small communities, all members are relied upon to sustain the community. Active participation by the users helps to keep the community active and everyone come together to solve the common problem.

It is often the case that there could be large numbers of non-participating users (lurkers) and small numbers of enthusiastic active users (posters). The posters are generally experts in the area and provide useful and insightful information. Lurkers use the information to learn and build their knowledge (Preece et al., 2004).

3.3.5 Short Lifespan

Any given PSN community will have a focused interest and a particular purpose. Most threads and post exist for a small period of time to discuss the problem. Once the problem is solved the posts are closed and the members dissipates.

The main posts in the community are ethereal and have a short lifespan. In some cases, they are persistent and provide information about the community. Membership in the community has various lengths of lifespan. Some people only come to have their problem solved. Other expert participants stay active for a long time.

3.3.6 Strong Incentive

PSN communities rely on users that actively participate in the community. The posts and threads are active for a small amount of time, so only interested people participate. Users focus on solving a particular problem and spend time and effort. Strong incentive are provided by the websites to motivate people to join and participate in the community.

The community would not sustain itself in the absence of sufficient motivation. The user needs motivation to contribute and if they receive proper incentive they will come back and contribute to the community. Some websites give contributors virtual points for their participation, some provide with badges and other incentives to keep them motivated.

These characteristics of StackOverflow and Reddit are studied in details in Chapter 5.

3.4 Benefits of Purposive Social Network

People are spending more and more time online. They connect with their friends and families, and do their work and shopping online. They use other users' recommendation

and experience to make decisions and solve their own problems. Social network analysis states the six-degree of separation theory (Milgram, 1967), each of us is connected to any random person in the world through the right six people we know. (Backstrom et al., 2012) in his paper showed the Facebook network has four degrees of separation. However, in the world of blogs, forums and messaging board this barrier is broken. One does not have to know anyone personally, professionally or at all to interact, comment or reply to one another (Monge and Contractor, 2003). This provides many benefits to the communities and for people to participate. The main motives for people to join such communities are discussed below.

3.4.1 Information Exchange and Self-interest

Online communities are a good place to share and exchange information because the Web is available to everyone, anywhere in the world. People come together and join a forum or a discussion board to share knowledge about their hobbies, health or politics. People use different websites to share different kind of information, Huffington Post ¹ to share the news to LiveJournal ² to blog about any topic.

In the area of E-Government, the government (e.g. USA has data.gov ³, UK has data.gov.uk ⁴) is publishing statistical and other Public Sector Information (or Open Government Data, OGD). People with different objectives come together and use this open government data and collaborate. They make apps to prevent crime using crime statistic of their geographic area or create support groups based on the health and morbidity statistic of their nearby hospitals. Websites like FixMyTransport ⁵, Patients-LikeMe ⁶, WhereDoesMyMoneyGo ⁷ are examples where people come together to utilize public data to solve social and political problems.

Social theorists consider self-interest as a major motivation for social actions. People maximize their social capital in the community that they later expect to profit from (Lin et al., 2001). People will contribute to the community for altruistic reasons as well as when they know they will gain something from it. People answers questions on StackOverflow to boost their points and improve their profile. This later lets them showcase their skills and find jobs.

¹www.huffingtonpost.com/

²www.livejournal.com/

³<http://www.data.gov/>

⁴<http://www.data.gov.uk/>

⁵www.fixmytransport.com/

⁶www.patientslikeme.com/

⁷wheredoesmymoneygo.org/

3.4.2 Symbiotic Relation and Social Exchange

People also communicate with their friends and family members using online communities. They may come together to form a neighbourhood watch program or start a political protest. People often go to forums to exchange recipes or to swap expensive tools they need for a short time. It is a mutually symbiotic relationship that both parties come out better off. BookMooch ⁸ and BarterPalace ⁹ web services are an example of this mutual beneficiary relationship between people in the online community (Ridings and Gefen, 2004). This mutual interest improves the quality of the communal knowledge.

In social networks, people often measure their importance by the number of friends they have or by the number of famous or important people are part of their network (Lin et al., 2001). People often also use the social networking platform for advertising their product and skills. The more people they have access to, the more benefit they get from the communities and their social network. Twitter and Facebook are good examples where people expand their social circle by connecting to people that increase their social capital (Wellman et al., 2001; Ellison et al., 2007).

3.4.3 Social Recognition and Personal Satisfaction

People use online communities to get information and share knowledge. Some people in the community have more experience and knowledge about certain topics. They are considered experts in the community. People use online communities to gain knowledge from the experts (Zhang et al., 2007). Some communities come together to create knowledge together for the benefit of the whole community. Wikipedia is a prime example of such communities. The expert users create most of the wiki articles and other users modify them and maintain the quality of the content. The person with the most knowledge and up-to-date information gets social recognition and is considered an expert (Huffaker, 2010).

People not only gain social capital and status with recognition, they also take pride and gain satisfaction from helping others. The respect they gain by showing and sharing their expertise with their peers sometimes makes them a power user. Online communities are a good place for people to build a reputation by sharing their expertise. It helps the community to achieve a goal, share information and motivate the users to participate actively (Anderson et al., 2012).

⁸bookmooch.com/

⁹www.barterpalace.com/

3.4.4 Recommendation System

People often leave reviews and recommendations of products in their blogs or forums on an E-business website like Amazon. This is also true for other recommendation websites for music, movies, restaurants, etc. People come together to rate and discuss the quality of places or products and also to get recommendation from each other. These websites also take advantage of the behaviour of the whole community to recommend resources to individual users based on their action (Angeletou et al., 2011). Recommendation systems take advantage of crowdsourcing and user generated data to ascertain the quality of product. They also collect data and create a user model and use it for marketing and targeting ads to users (Shang et al., 2011).

3.4.5 Expert Finder

People often go online to find a solution to their problems. An example of this is seen in the software developing community where people go to ask for help and solution to bugs in their program

People use specific forums for programmers, like StackOverflow, that have an extensive community of developers behind it. These communities help people with their code, bugs, system analysis, design, etc. Online communities are a good place to find experts, but it is also hard to find the right expert. Sometimes in popular websites questions gets buried below the amount of data that is created every day (Evans and Chi, 2008; Anderson et al., 2012).

These benefits help in developing PSNs and make it self-sustaining. Some of the benefits of StackOverflow and Reddit are studied in the Suman system.

3.5 Challenges in Purposive Social Network

PSNs provide a platform for people with similar interest and objective to come together and form communities. It is necessary to motivate people in the community to contribute and collaborate together to maintain a problem solving system. Websites using crowdsourcing, in some cases, are good examples of PSNs. Especially, where people are contributing and creating a knowledge base and utilizing the power of the network to achieve common goals.

PSNs are dependent on users' contribution and self-sustaining systems are difficult to model. An efficient PSN requires motivated users who actively contribute. Another important aspect is to find and retain users and keep them motivated with enough incentive to contribute is vital (Treude et al., 2011). These systems also use crowdsourcing

to manage and moderate the community and do quality control. The main challenges of building a strong PSN using crowdsourcing are discussed below.

3.5.1 Recruiting and Retaining Users

The three main parts of building a strong online community are starting the community, encouraging early interactions between members and moving to a self-sustaining interacting environment (Andrews, 2002). PSNs utilize users' participation and expertise to create and maintain a community. The question and answer forums like StackOverflow and collective knowledge generating systems like Wikipedia rely on a large amount of user participation to create an active community.

Generally, the users in a community can be distinguished as posters or lurkers. The posters are active users who frequently post and create content. The lurkers are passive users who do not contribute, but consume the contents. In the knowledge generating and question and answering communities the posters are the experts who generate the knowledge and lurkers are the ones who ask questions and learn from the knowledge (Nonnecke et al., 2006). There is some research done to motivate users to be more active and create content (Preece et al., 2004; Singh et al., 2009).

The community thrives when it reaches a critical mass and becomes self-sustaining (Raban et al., 2010). The primary focus to attract an initial base of users who will drive the growth is the quality of the contents. Another factor is providing the clear context within which the community could develop (Marathe, 1999). This can be achieved by having a focused interest and purpose.

There are several methods to get users to contribute like paying the users for their knowledge and expertise (Buhrmester et al., 2011). Another way is to make it a requirement for users to contribute as in reCAPTCHA where users have to digitize the image to finish the task. The popular option is asking for volunteers and giving them incentives to contribute (Zhou, 2011; Lampe et al., 2010).

One of the popular ways to attract new users is to make the platform free and easy to use. The system that targets a specific interest group with unique interest also gains new users by having a focused interest and use the users' network to expand (Ludford et al., 2004). These systems are open so people can contribute easily. Websites like Wikipedia, StackOverflow, YouTube are content specific and are free. People volunteer and contribute to these websites and create a vibrant community with like-minded people. The downside is that it is hard to predict how many people will actually participate and contribute in the whole process (Zhou, 2011). These systems require a good incentive model that keeps user motivated enough to contribute and maintain the quality of knowledge (Vassileva, 2012).

3.5.2 Incentive Model

Designing an incentive model for a PSN is complex. The system should make it easier for people to contribute, but also keep track of the quality of the content created. The game theory approach uses a proportional mechanism to rank the contents. Good quality content is rated higher and it is shown at the top of the page, hence it is rewarded by allowing more viewers. This generates more active participation and the low quality contributor is deterred by pushing the content at the bottom of the page. People who participate and generate high quality content are appropriately rewarded for their content as well as their good or bad behaviour (Ghosh and Hummel, 2011).

The game industry uses the instant gratification model to incentivize and motivate user for more participation and contribution. They make the whole process of creating content an enjoyable experience as a game playing scenario, in the case of ESP (von Ahn and Dabbish, 2004) and the user gets motivated to perform more task.

Question and answering systems rely on the users desire to be recognized. They provide different methods to measure and show a reputation or expertise in a particular field. When users establish reputation and is recognized as an expert in the area, the user generates more quality content and is motivated to participate in the community (Richardson and Domingos, 2003).

To sustain the constant quality of the content, there should be positive reinforcement for good quality contribution. Users who generate quality questions and promptly answer, they are rewarded for their contribution with badges or points. These users are motivated and are active in the community (Anderson et al., 2013). Similarly, when people are spamming or creating poor quality content or are asking repetitive questions should be given negative points and their contents should be eliminated for the lack of quality with the entry restrictions. Having high quality content brings back the users and they are more careful with the quality of their submissions (Ghosh and McAfee, 2011).

The other way to encourage user participation is to give the ownership of the content to its creator. This entitles the user and they become responsible with the maintenance and quality of the product. Also, creating a competitive environment where more contribution makes the user the top contributor of the category. This ensures a higher rate of returning and contribution of the participants (Singh et al., 2009).

An approval-voting scoring rule and a proportional-share scoring rule can enable the most efficient equilibrium with complements information, under certain conditions, by providing incentives for early responders as well as the user who submits the final answer (Jain et al., 2009). The approval-voting score rule is when the asker gives one point to each of the best answers. This helped in getting complement information and substitute information. The proportional-share scoring rule is when the asker gives some share

of the available points proportional to the user's answer. This helped in getting more substitute information.

3.5.3 Quality Control

Maintaining the quality of user generated content is often challenging in PSNs. High quality contents are necessary during the formative lifespan of the community to reach the critical mass. Good quality contents attract the initial base of the users and they attract their own network to join the community. The rate of information diffusion is higher in high quality content (Kwak et al., 2010). It helps in the growth of user base and helps in making community self-sufficient (Lin and Lee, 2006).

Many websites with a large amount of user-generated content depend on the joint community action to rank the content according to the quality by collective voting. This controls the quality of information displayed on the web page, the higher quality content is displayed more prominently and the lower quality content is surpassed and spams are removed. Users use thumbs-up or thumbs-down style ratings on questions on Stack-Overflow, reviews on Amazon, and posts on Reddit (Agichtein et al., 2008; Anderson et al., 2012).

These websites display higher quality contributions more prominently by placing them near the top of the page and pushing lower quality ones to the bottom. Since content displayed near the top of the page is more likely to be viewed by a user, ranking good content higher leads to a better user experience. Another benefit is that it also provides an incentive to produce high quality content that might appeal to a contributors' desire for attention (Jain and Parkes, 2009).

The rank order mechanism can be used to influence the quality of the content. Research has shown that the game theory model is used to motivate the attention driven users and generate higher quality content. This created a better environment for information distribution and sharing. The users that generate higher quality are featured prominently on the page and the proportional mechanism distributes the attention in proportion to the positive votes received. This creates a game theory equilibrium that facilitates higher quality posts and accordingly rewards the users creating a large incentive to participate in voting and contributing (Ghosh and Hummel, 2011).

A text analysis of the user content also determines the quality of the posts. A post with punctuation, grammar and typos can be analyzed to create an estimate of the knowledge and expertise of the contributor. Also, the syntactic and semantic complexity of the texts give an approximation of the overall knowledge of the user and their proficiency with the topic (Agichtein et al., 2008).

3.5.4 Search and Discovery of Quality Content

In the communities the amount of content created in user generated knowledge system can be huge. This makes it difficult to find the high quality content. There are many algorithms and models used to filter the best content from the masses and they are discussed below.

Link analysis and link based ranking is used in blogs and other social networking systems to form an estimate of the quality of the information. PageRank (Lawrence Page and Winograd, 1999) and HITS (Kleinberg, 1999) are the prominent ranking algorithms used in this method. The network graph formed by the analysis shows the flow of information and relationship between the people. Mutual reinforcement facilitates good blogs connecting to other good blogs and good questions getting good answers. A learning framework, that uses both relevance and quality for ranking, is efficient in getting the facts and trivia in a large-scale system (Bian et al., 2008). These can be used to filter high quality materials from the large collection of information available.

User voting and tagging is another use of crowdsourcing to search and discover appropriate information. Users vote the best questions and answers to the top of the web page and make it easier for people to discover the information. People also tag the content with appropriate keywords and categorize information that makes it easier to browse related content (Agichtein et al., 2008; Amitay et al., 2009).

User rating and recommendation are also used to search and discover high quality content as in the Amazon where books and other items are recommended based on user who bought an item also bought other items. Also, in IMDB ¹⁰ and Rotten Tomatoes ¹¹, movies recommended are given based on users' ratings and reviews. These website tracks users' popular behaviour and utilize the information to give recommendation to the rest of the population (Linden et al., 2003).

StackOverflow and Reddit use these aspects of crowdsourcing to maintain the contents and keep the community growing.

3.6 Case Study of Purposive Social Network

Earlier sections define and describe different types of PSNs. They provide insights into the challenges faced by them and different benefit of PSN when the community is growing and functioning. One of the research questions, KQ2, of this thesis is to study PSNs. This is done by studying StackOverflow and Reddit community data. In this section of

¹⁰www.imdb.com/

¹¹www.rottentomatoes.com/

the thesis StackOverflow and Reddit data are analyzed to get more information about PSNs.

The data of StackOverflow and Reddit are analysed using the empirical cycle in design science. They are studied in the context of PSNs.

The main research problem here is one of the research question KQ2. The sections below studies how PSNs are formed, how it is maintained and what incentivises the users to join and participate in the community. This is done by studying StackOverflow and Reddit community data. The main problems studied here are the social network, communication ties, user behaviour and how the community is moderated and quality of content is maintained.

The main research design here is aggregating and statistically analysing StackOverflow and Reddit data to understand user behaviour and incentive design. Communities ties and network structure are studied using social network analysis. All the data is empirical. They follow the standard research practice and are valid.

The main research execution and result analysis is discussed below for both StackOverflow and Reddit.

3.6.1 StackOverflow Analysis

StackOverflow website is well structured and provides lots of information about the questions, answers and users. It should be noted that the data is analyzed by myself using different statistical analysis tools. This is not done automatically by the Suman system. This analysis gives some insight into community size and users' activity.

Post Type	Number
Questions	7220639
Answers	12939535
Registered Users	3155396
Tags	36892
Unanswered Questions	2781360
Badges	9953324
Votes	63167067
Comments	30104484

Table 3.1: StackOverflow at glance as of June 2014

As of June 2014, there are over 3 million registered users in StackOverflow and more than 7.2 million questions asked by users. The questions are categorized using tags and individual users can subscribe to tags to receive daily, weekly or monthly email of all the questions asked in the tag. There are more than 36 thousand tags associated with various questions and answers. Users have casted more than 63 million votes to mark the quality of questions and answers.

3.6.1.1 Question and Answers

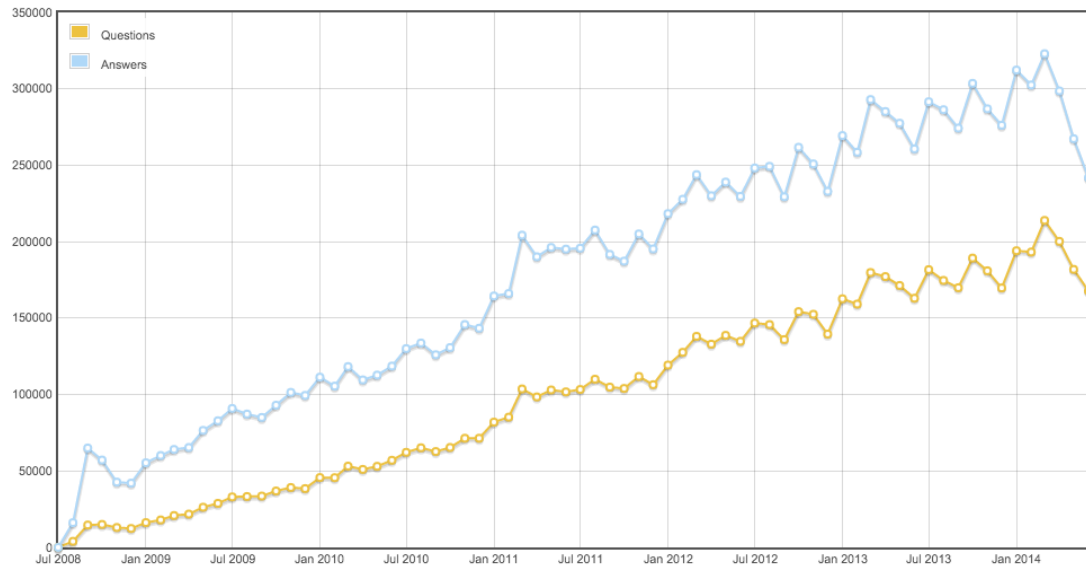


Figure 3.1: StackOverflow: Questions and answers posted every month

Figure 3.1 shows the number of questions asked and answered by users each month from the year 2008 to June, 2014. The trend shows the popularity of StackOverflow and user growth over time. On average, there are 1.03 million questions and 1.8 million answers posted each week.

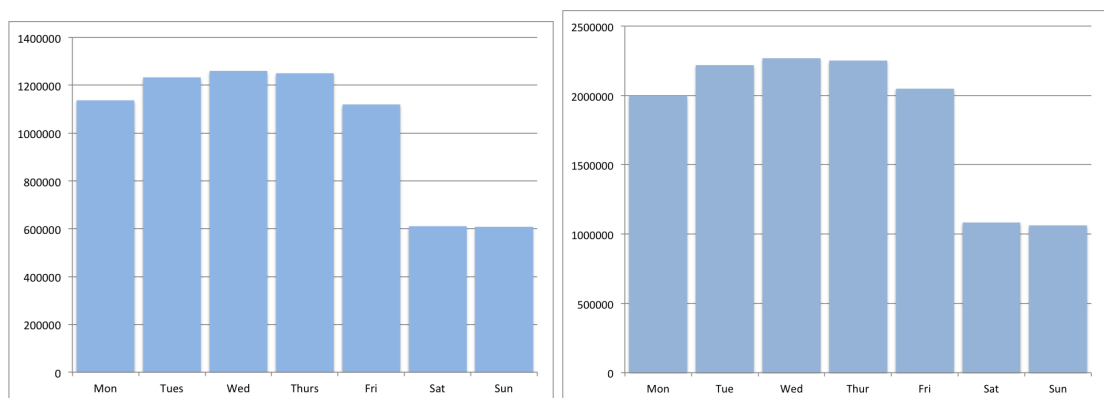


Figure 3.2: StackOverflow: Histogram of questions and answers posted the day of the week respectively

The analysis of posts shows that the programmers prefer to ask the questions and answers on weekdays as they are working on weekdays and find the problems in their program and post it on the website. Most questions and answers are posted on Wednesday.

Only looking at the answered questions, it's seen that on average a question receives 2.91 answers and 0.12% of questions receives more than 15 answers. So, the good questions do get a good response from the community.

The analysis of the unanswered question shows that despite high user feedback and participation 38% of the question in StackOverflow website has no accepted answers and 8% of the questions has no answer at all. These questions are either repeated questions or are too vague in description. When the question is not specific then the moderators flag it and users ignore the question.

Studying this trend of user activity gives insight into user behaviour and best practices to get the community engaged. This behaviour can be modelled and influenced to make more active and engaging community (Zhou, 2011).

3.6.1.2 Users

Users who contribute to the website are the main actors of these PSN. There are more than 3.1 million registered users in StackOverflow and they ask the questions, answer it, vote it and moderate the community. The users are not directly linked to each other to create relationships; in this network the relationship is formed by their interaction and their contribution. User behaviour, their motivation to use the website and incentive to contribute is described below.

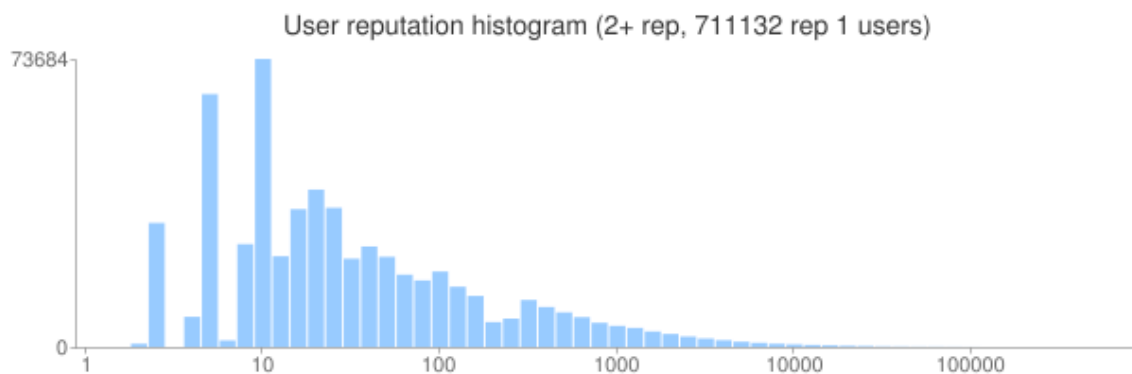


Figure 3.3: StackOverflow: User reputation histogram

Despite the high content generated by the users, 53.12% do not interact or contribute to the website, they have 1 reputation point that they receive while joining the website. The other problem is that new users are not familiar with the website etiquette and do not follow the community rules and it causes their questions to be ignored. When the user's activity is analyzed, it's seen that 19.74% of the users on the website asked only 1 question and never came back to the website and this percentage is even higher in the unanswered question. 17.8% of the unanswered question is made by the first time users and they didn't get any response so the user never came back. This shows that the users who do not get any response from the community do not return.

Understanding what makes users contribute to the community and what improves their experience will help in creating a more engaging and active community (Lampe et al., 2010).

3.6.1.3 Tags

In StackOverflow, the community is developed around similar interest and topics. This can be studied using the popularity of tags as well. The questions and the answers are provided with tags to categorize and arrange for easy search and discovery. The entire website is categorized using the tags and the list of most popular tags are shown in the table 3.2. Figure 3.4 shows the weekly use of the popular tags. The number of questions asked for each tag also provides an insight on the popular language used by developers at the time.

Tags	Number of Questions
C#	642126
Java	641667
Javascript	618902
PHP	578830
Android	497080
Jquery	481129
Python	307614
HTML	302990
C++	292515
MySQL	248507

Table 3.2: StackOverflow: Ten most popular tags and its instances

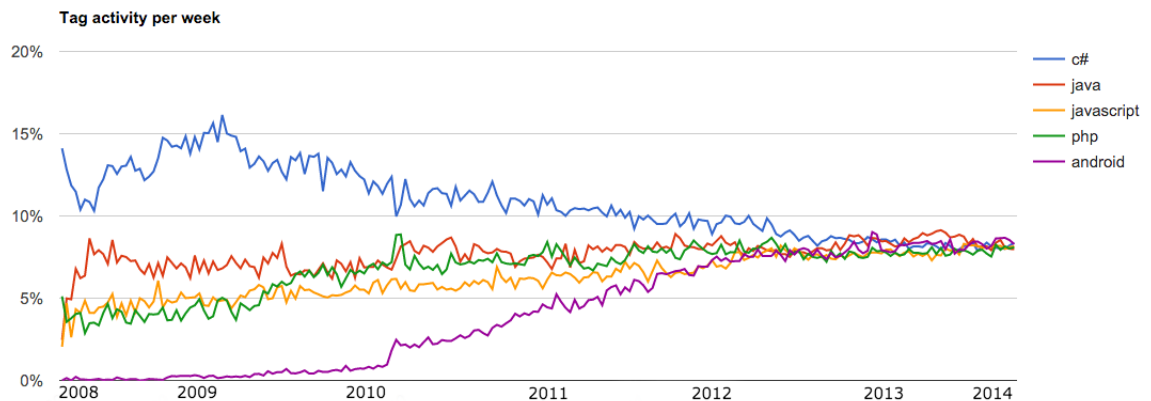


Figure 3.4: StackOverflow: Tag trends per week of 5 most popular tags

The analysis of questions shows that each question has between one to five tags associated with it. Most questions 54.7% have 3 to 4 tags associated with it. The relationship between the tags shows the overlapping of networks and how it is tied with one another discussed in details in the next section.

The other problem noticed in the unanswered questions is that they are not properly tagged. A quick analysis of the unanswered question tags shows that 35.2% of the unanswered questions have 1 or 2 tags. The difference in the tagging of the questions

could also cause in not getting an answer to the questions. The more tags a question has the more communities of people can view the question so the chances of it getting an answer increases. Also, the more specific tag a question has, the more specific experts can view the question and can give a detailed and higher quality answer. The new user makes this mistake and lose out on a quality users and expert audience. The details of the relation on number of tags and how it differs in answer and unanswered question can be seen in the table below.

Number of Tags used	Unanswered Questions (%)
1	11
2	24
3	29
4	21
5	14

Table 3.3: StackOverflow: Percentage of unanswered questions with number of tags

Table 3.3 shows the problem with the relation of the number of tags and how it affects answering a question. This doesn't explain the problem if the question has wrong tags. In this case, the question reached the wrong community of users and it was ignored because it wasn't in their area of expertise.

During the initial analysis of the data, it was attempted to figure out if users used the wrong tags for their question. There was no accurate way to answer this question. A simpler solution was used to figure out this problem. First, a quick comparison was done to find the number of unanswered questions from big and general tags like Java, Python, etc. with the specific versions of the tags like Java-EE, Python 3.1, etc.

The analysis showed that the percentage of unanswered question was lower in the specific tags than the general tags for 1000 questions. Table 3.4 shows a detailed description of unanswered questions in the top 3 tags and its specialized tags.

This doesn't really provide a concrete answer to the question as correlation doesn't mean causation, but it does give some insight into the problem. The problem could be that the general tags get a lot more question each day so the questions can get easily buried deep and couldn't reach the right people. (Jain and Parkes, 2009) shows that in these forums most users usually read the top page and the view count has a steep decline as we go further into the next pages.

The other problem could be that the big communities are heavily moderated and if the moderator sees any problem with the post, they quickly remove it. Also, these pages have a lot more moderator and each moderator have to stick by the community rules, but they have their own views as well. The analysis shows that the percentage of removed

Tag	Question count	Unanswered Question	Unanswered question (%)
Java	641667	42136	6.566
Java-7	1596	149	9.33
Java-EE	17903	1497	8.36
Javabeans	1968	171	8.68
Android	497080	55037	11.07
Android-4.0	1206	163	13.5
Apk	2222	241	10.84
Adb	2040	211	10.34
Python	307614	16698	5.42
Python-2.7	19264	1549	8.04
Python-3.x	12499	578	4.62
Python-idle	579	44	7.59

Table 3.4: StackOverflow: Difference percentage of unanswered questions in general tags vs specialized tags

questions by moderators are higher in the general tags 2.85% than in the smaller tags 0.9%.

Another view is that the users who answer the questions on general tag pages find the best question suited for them in the big heap of questions, they don't see the incentive to look through each question and try to answer as much as possible because of limited time.

3.6.1.4 Votes and Reputation

The other reason more questions remained unanswered in popular tags is because users can get more votes by answering questions in the bigger tags then the smaller tags. The analysis of votes shows that in the table below. Here general tag is considered top 10 tags with more than 100k subscribers and specialized tags are the smaller tags of the top 10 general tags with less than 20 thousand subscribers.

Answer votes	General Tags % (JAVA)	Specialized Tags % (JAVA EE)
0	25.15%	18.1%
1	29.31%	36.78%
>1 - <=3	27.02%	28.92%
>3 - <=5	9.5%	8.68%
>5 - <=10	4.04%	4.72%
>10	3.96%	1.94%
<0	1.02%	0.86%

Table 3.5: StackOverflow: Percentage of Answers' vote count in general tags vs specialized tags

This analysis does not confirm the quality of the answers, just the votes. It can't be said that the quality of answers are better in bigger tags than in smaller tags. It could be just because of the number of users in each tag and how many of them vote for the questions and answers. The users' behaviour changes based on the size of the community (Zhou, 2011).

3.6.1.5 Communication network structure

StackOverflow is not a typical social networking website per se as users cannot create explicit friendship or follow other peoples' work, they cannot send private messages or form groups.

On StackOverflow, social ties are formed implicitly by users asking questions and answering them. These social ties are also formed by voting on the posts and commenting on them. The communication network is studied to see the social ties of the individuals (Monge and Contractor, 2003).

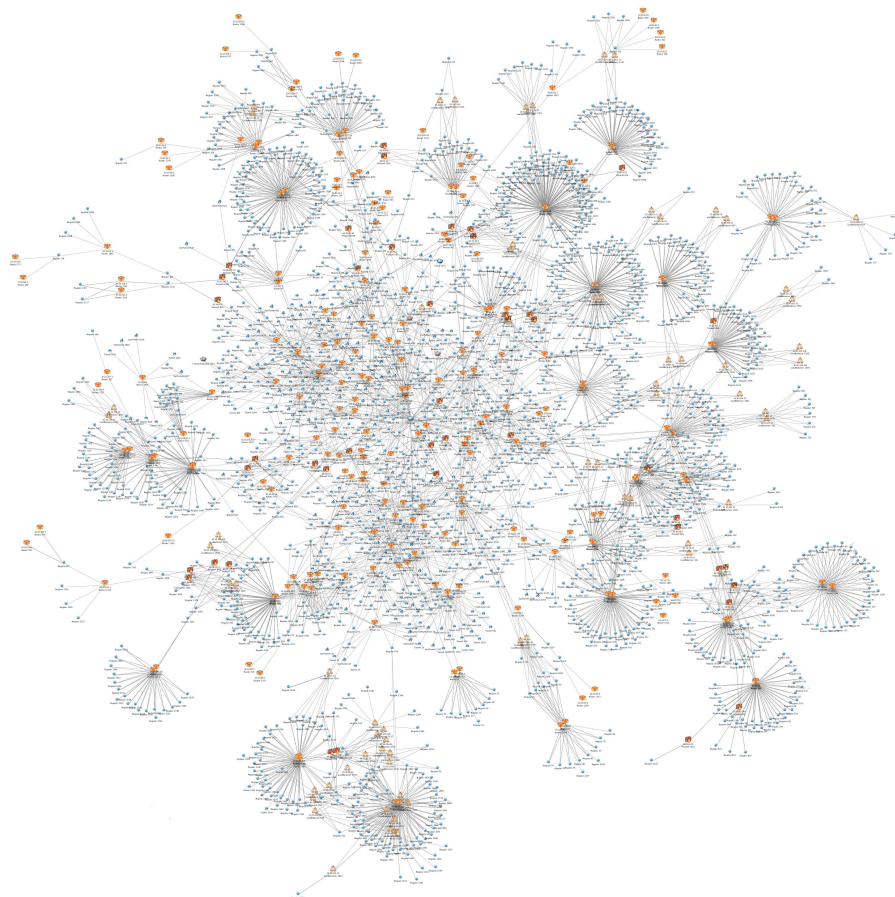


Figure 3.5: StackOverflow network structure

For the network analysis of StackOverflow dataset, the communications between users are studied. Here, the users who ask questions, give answers, create posts and comments are nodes and edges are formed between the users who interact with them. The relationship between users becomes stronger the more they interact with each other.

Figure 3.5 shows the sparse and interconnected communication network build over time. The figure 3.5 shows the top 20 tags and how the community is interconnected. The tags that are used together are strongly connected and overlap. The tags that are rarely used with popular tags stay at the edge of the network.

The nodes on the edge of the graph are a small niche community that doesn't overlap with popular tags but are sometimes linked together by one or two questions. The users in the middle are heavily connected to a lot of other people. They are the expert users who answer a lot of questions and are at the center of the network. There are also expert users who connect one network to another, they have knowledge of different programming languages and connect two or more communities together.

In the network analysis, the edges get stronger the more users interact with each other, but closer analysis of the network graph shows that the edges are quite weak between the nodes. Only few nodes have medium strength nodes.

3.6.1.6 Tags network

In social network analysis, community is detected by analyzing the linkage of the network. Another approach to detect community is analyzing the social objects and detect the topical clusters. This gives a network structure based on similarity of topics (Zhao et al., 2012). Similar approach is used for community detection of StackOverflow. To analyze the community structure interconnectedness of tags are studied.

In StackOverflow questions have multiple tags attached to each other. These questions appear in all the different tags page and people subscribed to each tags can view and answer the questions. This makes the communities overlap with each other and show the strength of the relationship between the communities.

The figure 3.6 shows the relationships between the most popular tags and how closely they are related to each other. In the following graph each segment size is directly proportional to the number of instances it is used and the connection between the tags indicate the times they have been used together in a question. The thickness of the connection shows the strength of the relations. The segment is colour coded by the frequency of connections, red segments are strongly connected and blue segments are weakly connected.

The clustering of the tags shows the relationship between the tags and technologies. The two popular tags, JAVA and Android are closely related to each other, but are scarcely

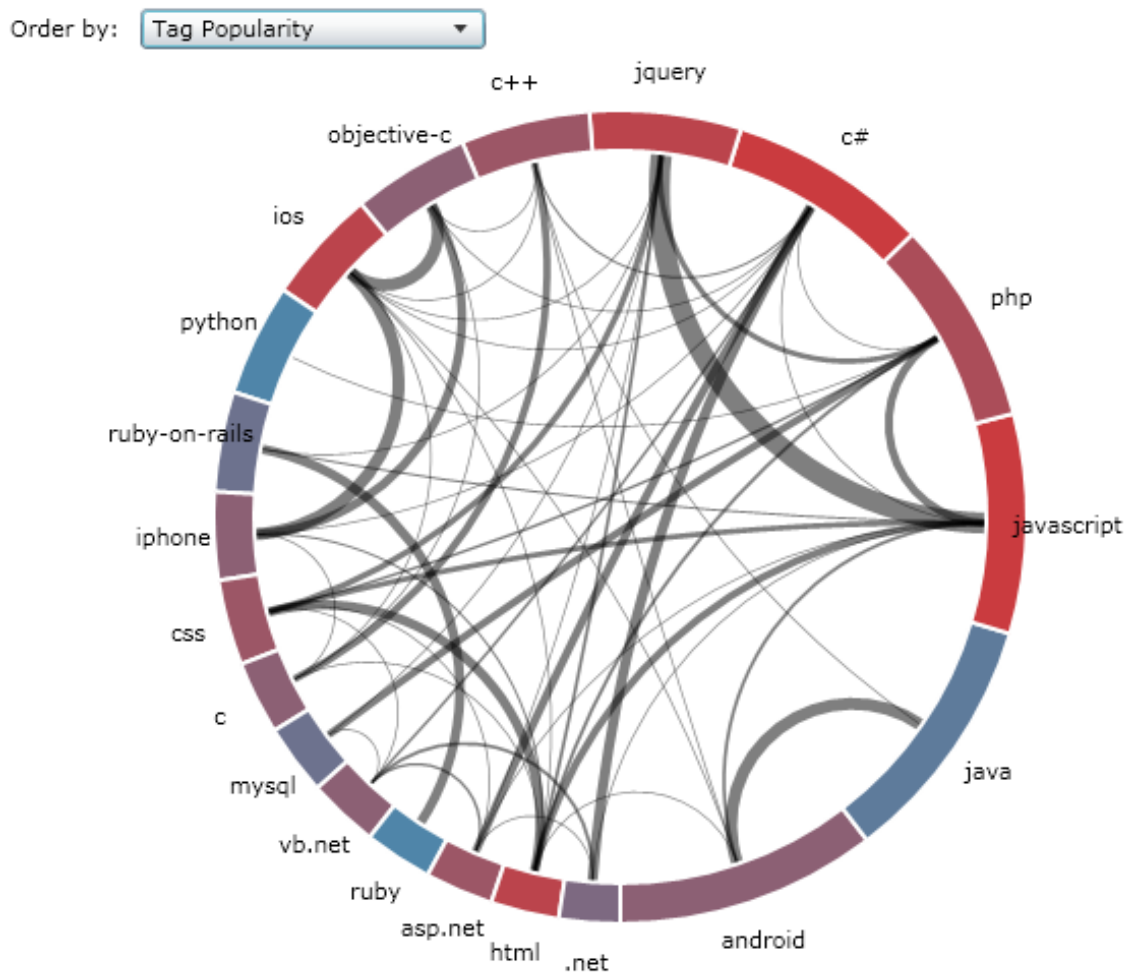


Figure 3.6: StackOverflow: Popular tags clustered together

joined with other tags. The strongest relationship is between jQuery and JavaScript because the overlapping framework of the two programming languages. C, C++ and C# are also a closely related groups as well as iOS, Objective-C and iPhone. However, sometimes Objective-C is also tagged with C, C++ and C#.

There is a large cluster of connected web development languages, CSS, HTML, JavaScript, and jQuery, indicating the close knit use of these technologies in development of website and web applications. The interesting thing is the relationship between the scripting language PHP and Python, they are popular tags but are sparsely connected with other tags and are weakly linked with database related tags.

3.6.1.7 Incentive Design

StackOverflow is one of the most popular Q&A websites for software developers. The Alexa's ¹² rank of StackOverflow is 57 (Alexa, 2015b) worldwide as of November, 2015.

¹²<http://www.alexa.com/>

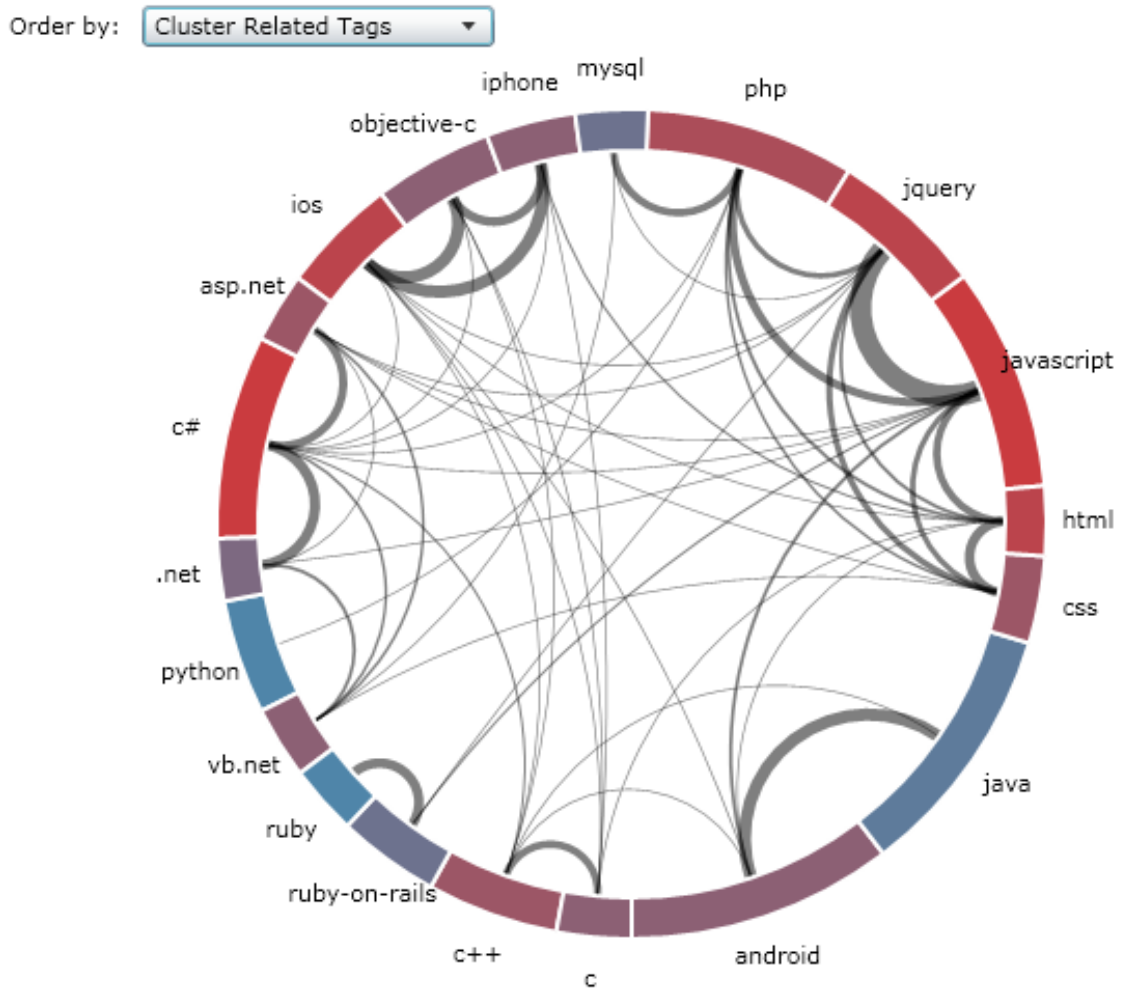


Figure 3.7: StackOverflow: Related tags clustered together

StackOverflow have a strong incentive design to motivate people to contribute and grow the community. They maintain high quality content by crowdsourcing that leads to the popularity of the website. PSNs require active user participation and this requires strong incentives for users to contribute (Ghosh and McAfee, 2011).

StackOverflow uses a game theory model to encourage user participation and activity. Participation is encouraged through an elaborate point system and users also receive badges for participation. The top contributor and user with highest reputation are featured on the question page, giving the user more visibility and acknowledgement of the user's expertise. This encourages participants to accumulate more points and contribute to get recognition (Anderson et al., 2012).

When an answer is voted up, the user gains 10 reputations and 5 points when the question is voted up. When an answer is accepted the user receives 15 points and there is also a negative point system, a user loses 2 reputation points when a question or an answer is voted down (StackOverflow, 2008). This keeps the spamming in check and repeated questions and answers are avoided.

Reputation	Number of users
1	1666503
2-10	327106
11-100	767635
101-1000	302338
1001-10,000	78766
10,001-50,000	7467
50,001-100,000	650
100,001-200,000	238
>200000	91

Table 3.6: StackOverflow: Number of users with reputations

As table 3.6 shows, there are 1666503 users (52.81%) with 1 reputation point and one user with 827131 reputation points. The distribution of the users' reputations shows that more than half of the users are lurkers and the elite users with the most reputation points are the editors and moderators of the community and are considered the expert in their field.

The reputation of the user has a direct correlation with the trust in the community. StackOverflow has designed an excellent reward program to motivate and incentivize the users to contribute and gain more reputations and badges.

The system also encourages users to participate as the higher reputation points gain more privileges. When a user has 15 reputation points, only then they can up vote and 50 points allow users to comment. To stop harassment and spam, user requires 125 reputation points to vote down and it costs the user 1 reputation point. The incentive model is thorough and higher reputation points open more gates for users to interact and contribute and be acknowledged as the expert in their field.

Currently, there are 77 different types of badges given to the user based on their contribution. There are badges given to the user who asks questions with 1 reputation point (Student), to the user who edits the answers to make the posts better (Editor) and to even an active user for a year (Yearling). This type of virtual acknowledgement of efforts encourages the user to participate and contribute to the website.

The other method that encourages the users to participate is the promptness of the response. The asker prefers to receive information sooner rather than later, and will stop the process when satisfied with the cumulative value of the posted information. The analysis of the posts shows that half of the questions get an answer within an hour of the posting and within a day the questions receive an accepted answer.

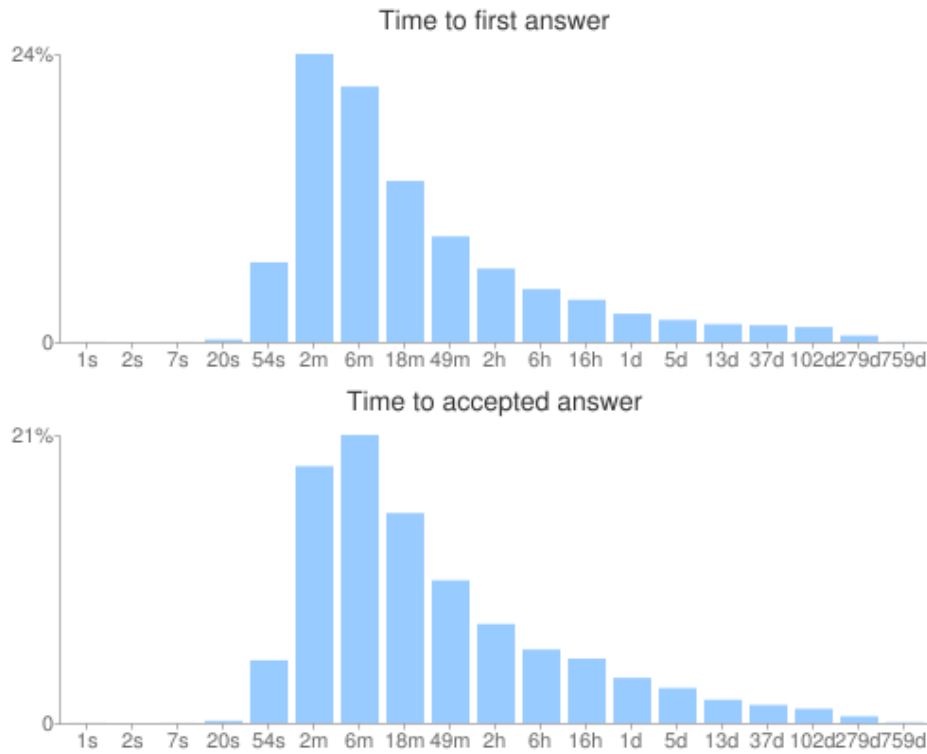


Figure 3.8: StackOverflow: Time histogram of questions receiving first answer and accepted answer respectively

3.6.1.8 Quality Control

The community thrives because of the high quality of content and it is possible by the user's action and moderation. Users vote up the best questions and answers and vote down the bad quality content or repeated posts. There are more than 6 million votes casted in StackOverflow and the user with enough reputations are allowed to cast 40 votes per day. Reddit doesn't have any limit in casting votes, there are more than 46 million votes in the collected dataset.

Vote	Question Vote Count	Answer Vote Count
0	5026503	6692276
1-10	4749669	12239034
11-100	246576	515107
101-1000	1890	26909
>1000	260	489
<0	608453	284509

Table 3.7: StackOverflow: Questions and answers vote count

The analysis of StackOverflow questions and votes shows that every question receives 3.06 votes on average. One question received negative 145 votes and the highest vote received to a question is 12822. Similar analyses of answers and their votes shows that, on average an answer receives 0.99 votes and the lowest vote to an answer is negative

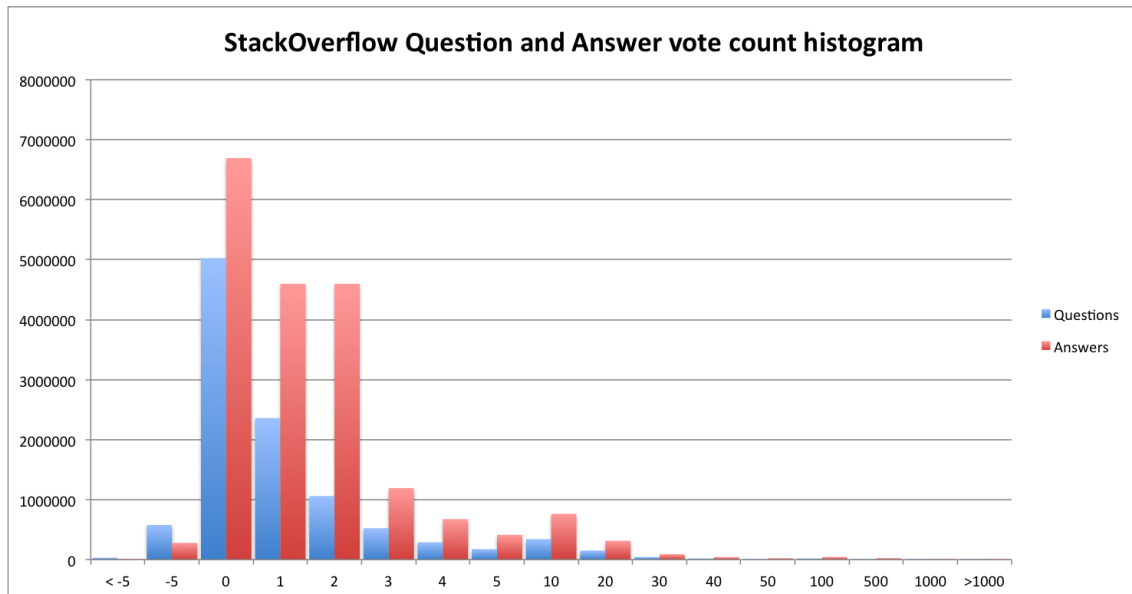


Figure 3.9: StackOverflow: Questions and Answers Vote count histogram

54 and the highest vote is 18215. There are more questions (69.61%) with 0 votes than answers (51.71%).

3.6.1.9 Community Moderation

Another way the quality of the community is maintained is through moderation. The community members elect moderators and they are considered expert in the field. They are given responsibility to edit and remove the posts. They keep the spamming in check and maintain the quality of the posts in the community. They also have the authority to ban users if they are found misbehaving and spamming.

StackOverflow provides additional badges for moderators. They also have extra privileges than the normal user and have sometimes access to the community interactions and extra information.

There are 18 moderators in StackOverflow and they are elected democratically by popular election. They moderate (edit and remove) 1500 flagged posts daily on average.

3.6.2 Reddit Analysis

The structure of Reddit is completely different from StackOverflow, hence the same analysis could not be done with the Reddit dataset. First, the Reddit posts do not have any tags, so they couldn't appear to multiple communities of users. The posts are posted in a particular subreddit and they couldn't be shared across different subreddits. To do so, user had to create a brand new post with the same question on a different subreddit.

Post Type	Number
Posts	19502
Comments	414591
Registered Users	71079
Votes	46112303

Table 3.8: General overview of Reddit dataset

Reddit posts are threaded, a parent comment has multiple child comments and users mostly deflect from the main topic and start a discussion on a different topic, this sometimes is also personal. This is completely different from StackOverflow because users use answer posts for answers and any other conversation they want to have is strictly restricted to the comments section. This caused a major problem with the Reddit dataset, as it couldn't be decided which comment or the child comment is the actual answer to the question. Further on, unlike StackOverflow, there is no 'Accepted Answer' tag for Reddit posts. So, it was quite hard to understand if the question got a valid answer or not.

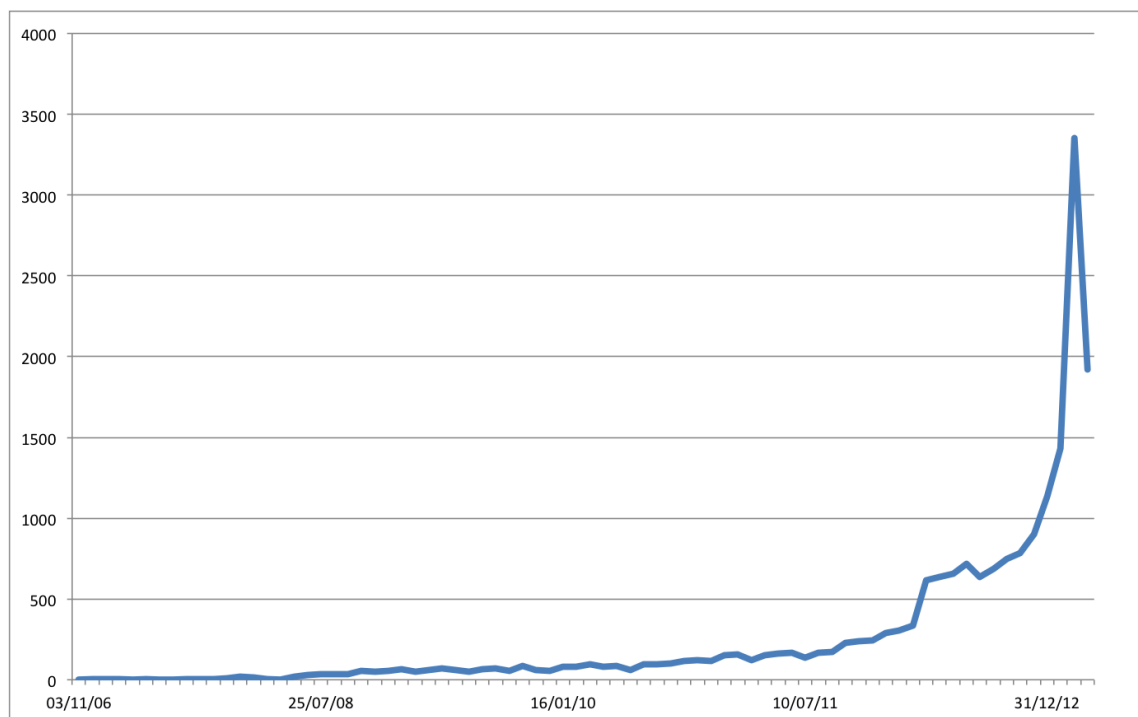


Figure 3.10: Reddit: Posts created per month

The analysis of 11 different technical subreddits, 19 thousand questions and 0.41 million comments is done quite differently and separately from StackOverflow datasets. A post has 20 comments and 140 votes on average. A comment, however, has 11 average vote, and 3.5 child comment.

Here, if any post has no comments with a positive vote or no comments at all, then it is considered an unanswered question. Any post that has at least one positive comment,

for the purpose of this analysis, it is assumed that it answers the question. This is not a good strategy as it has lots of uncertainty, but no other way had been found to solve the problem. Only one subreddit ‘/r/learnprogramming’ has a ‘Solved’ flair that gives a confirmation that the question has been answered.

The posts on these subreddits are not only questions but also discussion posts and news posts where users are sharing links to other websites and blogs. There are no ways to distinguish between these types of posts. So for the analysis, any post that is linked to the external source was considered a discussion post or news post. In the big subreddits like /r/javascript, /r/python, 79% and 81% of the posts are discussion and news posts respectively. They are linked with an external website. Also, 2.03% are the cross post, these posts are Reddit post linked from another subreddit. The number of discussion and news posts drop significantly in the ‘learn programming’ subreddits. Only 4.8% of the posts are linked post, 4.6% are cross posts and the rest are question posts. However, the ‘programming’ subreddit, 99.7% of posts are discussion and news posts.

Type of posts	Number	Percentage
Question post	6367	32.7%
Discussion post	13115	67.2%
Posts with ‘Solved’ flair (learnprogramming)	105	4.05%
Parent comments	88881	22.5%
Child comments	305710	77.4%
Posts with 0 vote	1687	8.65%
Posts with -ve votes	1059	5.43%
Posts with vote >5	14496	74.3%
Posts with no comments	3952	20.26%

Table 3.9: Reddit: Different types of posts

There are no tags on the subreddit posts, so the name of the subreddit is added as a tag to the post. Sometimes in the general subreddit like /r/learnprogramming users add the name of the programming language in the square brackets, e.g. [python]. These names are also added as tags to the posts.

Using all the generated data and using the subset that includes the question data, the dataset has been analyzed. 20.26% of posts have no comments at all and it is assumed as unanswered questions. 57.8% of these unanswered questions have 0 or negative votes. It is supposed that the reason the posts have no votes is because users have ignored them.

The percentage of unanswered questions created by a new user with no votes are 39.1%. It is much easier to create a new account in Reddit as it doesn’t require an email address, only a username and password. So it is assumed users create a new account, just to ask a question and never use it again.

There are very few 2.8% unanswered questions that have more than 5 votes, but no comments at all. So it is assumed that if the question is interesting, the users engage in some kind of discussion. The majority of answered posts 78.63% have more than one comment. The table below gives a general overview of all the analysis done.

Since the Reddit API doesn't allow getting more than 2000 posts for any subreddit, all the posts are not collected. So then trend and analysis is done only on extracted data. So a long term posting trend cannot be done on the Reddit data, but using a week's data is clear that users post most on Wednesday.

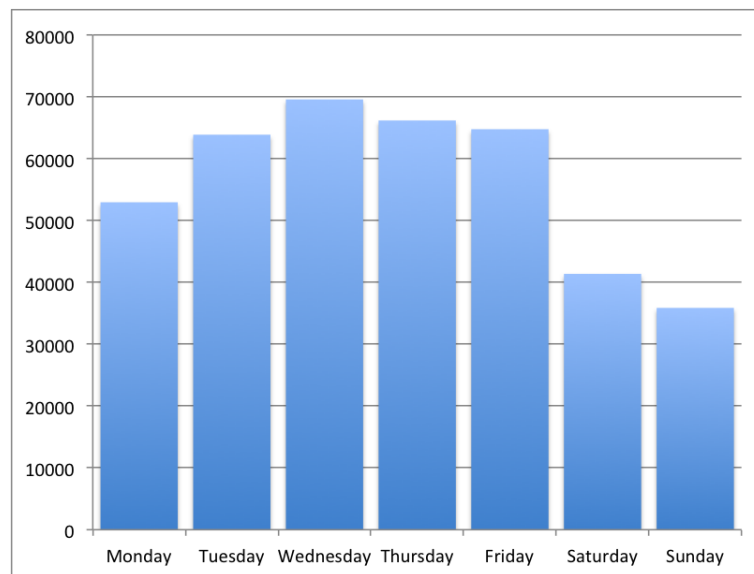


Figure 3.11: Reddit: Histogram of posts created day of the week

3.6.2.1 Communication network structure

Reddit allows user to make friends and send private messages but this information are not public and cannot be downloaded by API without user's explicit consent. So the social ties are formed implicitly by users asking questions and answering them. These social ties are also formed by voting on the posts and commenting on them. The communication network is studied to see the social ties of the individuals (Monge and Contractor, 2003).

For the network analysis of Reddit dataset, the communications between users are studied. Here, the users who ask questions, give answers, create posts and comments are nodes and edges are formed between the users who interact with them. The relationship between users becomes stronger the more they interact with each other.

Similar things as StackOverflow can be observed by studying the communication network of Reddit. The nodes in the center are connected to multiple nodes and acts as bridges between different communities. The Reddit data are incomplete, so the network is not

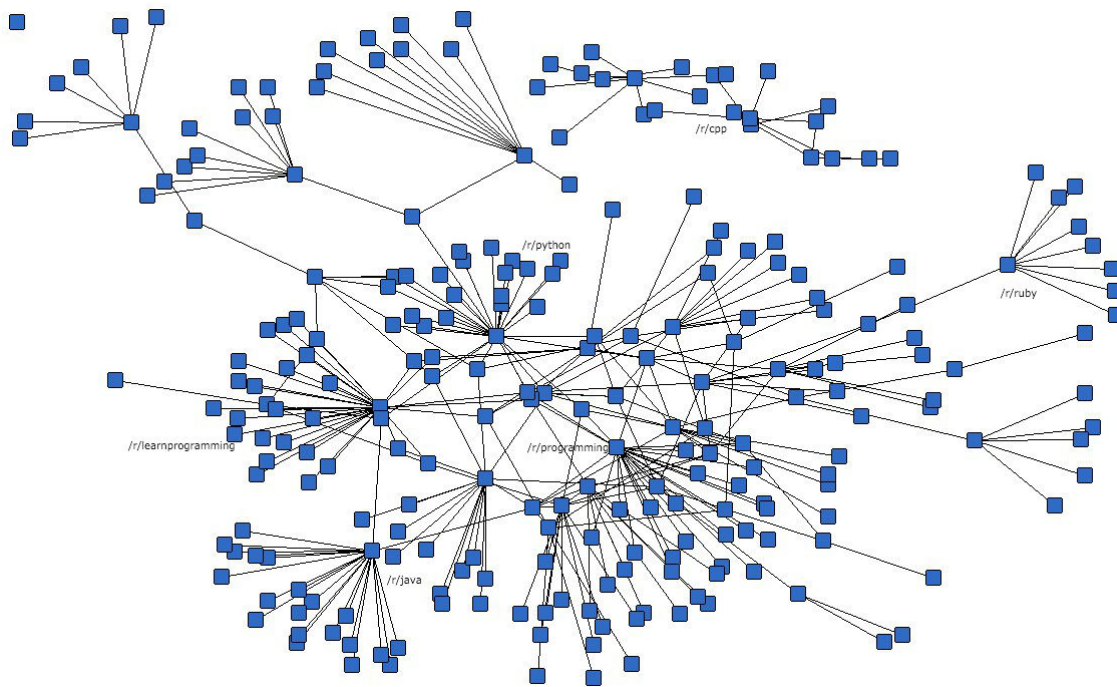


Figure 3.12: Reddit network structure

connected properly. The communities are linked by weak ties where a post is created across subreddits.

The only thing different here is that the edges are thicker in the Reddit community than StackOverflow. This could be because Reddit has a relaxed commenting rule where users can have informal discussions. StackOverflow on the other hand has strict community guidelines where users can only interact with comments on the question or answers.

3.6.2.2 Subreddit network

In Reddit, the cross posts are inspected to find the relationship between the subreddits because the Reddit data does not have tags. There were only 254 cross posts in the dataset, so the relationship between the subreddits is not as strong as StackOverflow but it follows the similar pattern.

Looking at the users' interaction also showed the user behaviour in subscribing to a tag or subreddit.

3.6.2.3 Incentive Design

Reddit is one of the popular websites for people to share web links with people and it's also popular with programmers but not as much as StackOverflow. The Alexa's ¹³ rank

¹³<http://www.alexa.com/>

Tag/Subreddit	Reddit Users	StackOverflow Users
Python	101893	80400
Javascript	74768	134600
Java	43875	124800
PHP	38661	84600
C++	31654	64200
C#	2507	85300

Table 3.10: Number of subscribed users for each tag and subreddit in StackOverflow and Reddit respectively

of Reddit is 32 (Alexa, 2015a) worldwide as of November, 2015.

Reddit have a strong incentive design to motivate people to contribute and grow the community. They maintain high quality content by crowdsourcing that leads to the popularity of the website. PSNs require active user participation and this requires strong incentives for users to contribute (Ghosh and McAfee, 2011).

Reddit follows the same model, here the reputation points is called ‘karma’ and a user can have link karma from creating posts and comment karma by making comments. The karma is the accumulation of all the up votes and down votes. But the link karma and comment karma are not added together. Higher the karma points, more popular their posts are and Reddit algorithm puts their post at the top of the page.

Karma	Number of users with link Karma	Number of users with comment Karma
0	565	9994
1	862	38156
2-10	3858	88237
11-100	8709	36222
101-1000	1963	2497
>1000	287	11
<0	984	16932

Table 3.11: Reddit: Number of users with karma

As table 3.11 shows, there are 862 users (1.21%) with 1 link karma and 38156 users (53.68%) with 1 comment karma. There is one user with 2629 comment karma and one user with 4077 link karma. The distribution of the users’ reputations shows that more than half of the users are lurkers and the elite users with the most links and comment karma are the editors and moderators of the community and are considered the expert in their field.

Similar to StackOverflow badges, Reddit has an award system when a user can earn different awards based on their contribution and it is visible in their trophy case. Currently, Reddit offers 34 awards, they can be as simple as ‘Verified Email Address’ given to users who verify their email address to ‘Best Link’ or ‘Best Comment’ when they

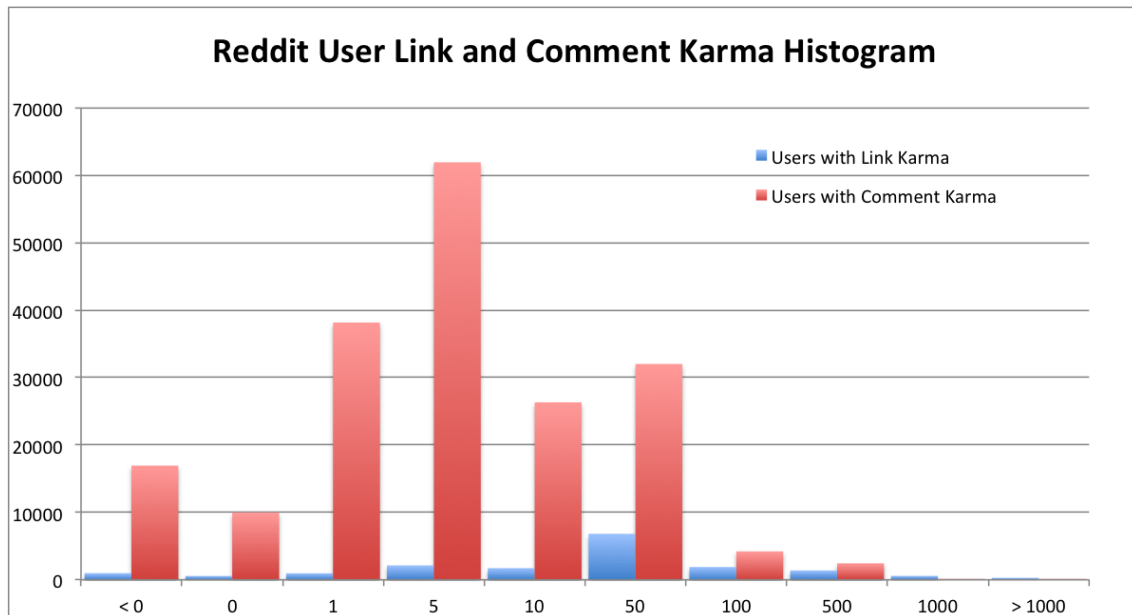


Figure 3.13: Reddit users' link and comment karma histogram

contribute positively to the community. The response time in Reddit is quite fast. All the posts that receive any comments get the first comment within two hours. Half of them receives the first comment within half an hour. This also shows that the lifespan of posts are not long, if the post doesn't get a comment within 3 hours, it doesn't get any comments.

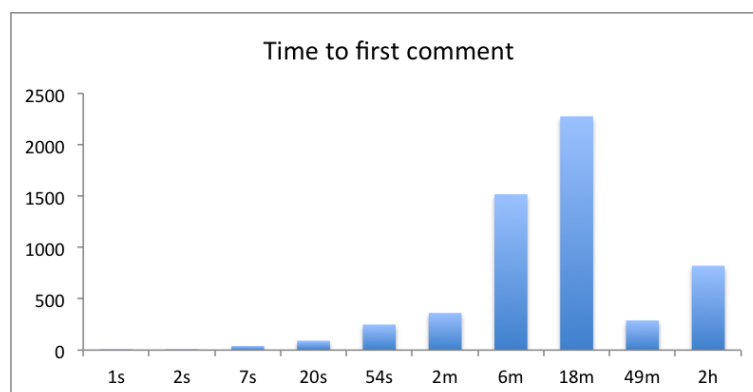


Figure 3.14: Reddit: Time histogram of posts receiving first comment

Another good incentive model Reddit uses is the award of 'Reddit Gold'. Sometimes users treat 'Reddit Gold' as an actual award to an excellent response or offer as a reward for certain type of contribution. Gold is rewarded by one user to another and it provides special privileges to the Gold winner. They get an ad free and custom experience of the website and get access to a special 'Gold' subreddits.

This kind of reward system helps get prompt responses and get good quality contributions. The analysis of the posts shows that half of the posts get a comment within an hour of the posting.

3.6.2.4 Quality Control

The community thrives because of the high quality of content and it is possible by the user's action and moderation. Users vote up the best questions and answers and vote down the bad quality content or repeated posts. Reddit doesn't have any limit in casting votes, there are more than 46 million votes in the collected dataset.

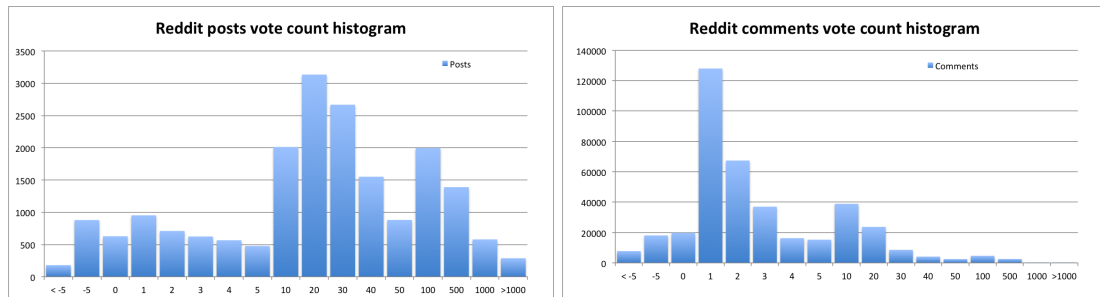


Figure 3.15: Reddit: Posts and Comments Vote count histogram

Reddit analysis shows that questions receive 35.87 votes on average. One post received negative 21 votes and the highest vote received to a question is 4077. Similar analyses of comments and their votes shows that, on average a comment receives 0.35 votes and the lowest vote to a comment is negative 266 and the highest vote is 2629. Same as StackOverflow, there are more Reddit comments (4.78%) with 0 votes than posts (3.22%). But the total percentage of posts and comments are not as much as StackOverflow.

Vote	Post Vote Count	Comment Vote Count
0	628	19821
1-10	5341	303042
11-100	10229	43696
101-1000	1968	2641
>1000	287	11
<0	1059	25880

Table 3.12: Reddit: Posts and comments vote count

3.6.2.5 Community Moderation

Another way the quality of the community is maintained is through moderation. The community members elect moderators and they are considered expert in the field. They are given responsibility to edit and remove the posts. They keep the spamming in check and maintain the quality of the posts in the community. They also have the authority to ban users if they are found misbehaving and spamming.

Reddit provide additional badges for moderators. They also have extra privileges than the normal user and have sometimes access to the community interactions and extra

information. In Reddit, average number of moderators in a subreddit are 6 and no data is freely available about their moderation activity in the community.

3.6.2.6 Discourse Analysis

All the statistical analysis done of the StackOverflow and Reddit data are done by analyzing the data using SPSS. A discourse analysis was also done by looking through the data to find interesting posts and general trend that is not properly addressed by statistical analysis applications.

Looking through the questions and answers it is evident that StackOverflow interactions are more formal and structured than the Reddit interactions. Reddit post are more discussion oriented and people share current news and updates. StackOverflow questions are more problem oriented questions and sometimes there are also discussions, but they occur mostly in the comments and chats.

On average, StackOverflow questions and answers are longer than Reddit questions and comments. Also, Reddit posts get more votes than StackOverflow posts. A quick analysis of one thousand posts from each website also shows that there are more spelling errors in Reddit posts than StackOverflow posts.



Figure 3.16: An example of Reddit post (programmer564698, 2015) where people share personal stories

The posts and comments on Reddit are more personal. Users ask more indirect questions like asking for recommendation for books, tools, etc. User also shares lots of memes and funny posts that are not clearly related to the programming language. User's also share personal stories, anecdotes and experience.

What's the difference between JavaScript and Java?


78

★


225

share

edited Sep 4 '10 at 19:29

 Peter Mortensen
9,109 ● 10 ● 63 ● 98

asked Oct 28 '08 at 22:10

 Guy
21k ● 63 ● 182 ● 254

575

✓

share

answered Oct 28 '08 at 22:12


 Greg Hewgill
451k ● 98 ● 797 ● 992

Figure 3.17: Example of funny StackOverflow response (Guy, 2008)

However, in StackOverflow, most of the posts are about solving technical problems, but there are a few posts where people share jokes, humorous anecdotes and trivia. There was a question asking users to share jokes that have 459 answers¹⁴ and one where users are sharing their favourite "programmer" cartoon¹⁵ that has 135 responses.

HTML-plus-regexp will liquify the nerves of the sentient whilst you observe, your psyche withering in the onslaught of horror. Regēx-based HTML parsers are the cancer that is killing StackOverflow *it is too late it is too late we cannot be saved* the transgression of a child ensures regex will consume all living tissue (except for HTML which it cannot, as previously prophesied) *dear lord help us how can anyone survive this scourge* using regex to parse HTML has doomed humanity to an eternity of dread torture and security holes *using regex* as a tool to process HTML establishes a breach *between this world* and the dread realm of corrupt entities (like SGML entities, but *more corrupt*) a mere glimpse of the world of **regex parsers for HTML will instantly** transport a programmer's consciousness into a world of ceaseless screaming, he comes, the pestilent slithy regex-infection will **devour your** HTML parser, application and existence for all time like Visual Basic only worse *he comes he comes do not fight he comes, his unholy radiance destroying all enlightenment,* HTML tags **leaking from your eyes like liquid** pain, the song of regular expression parsing will extinguish the voices of mortal **man from the sphere** I can see it can you see *if it is beautiful the final snuffing of the lies of Man ALL IS LOST ALL IS LOST* the pony he comes he comes he comes the ichor permeates all MY FACE MY FACE oh god **no NO NOOOO NO** stop the angles are not real **ZALGO IS TONY THE PONY, HE COMES**

Have you tried using an XML parser instead?

Figure 3.18: Example of funny StackOverflow response (Jeff, 2012)

There are some questions that have funny and controversial answers too. A question from a user about XHTML that gave the humorous response showing importance of space.

¹⁴<http://stackoverflow.com/questions/234075/>

¹⁵<http://stackoverflow.com/questions/84556/>

There are controversial questions where people share their personal experience that get so many responses that moderators have to close the thread.

What's your most controversial programming opinion?

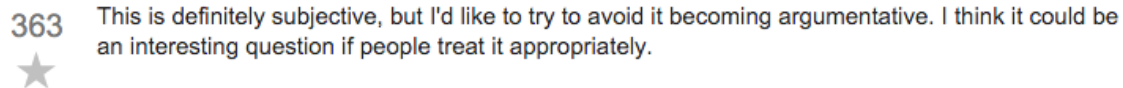


Figure 3.19: Example of StackOverflow question (Skeet, 2009) where people are sharing their opinion

People also use this platform to share trivia about different languages and their programming experience.

Strangest language feature

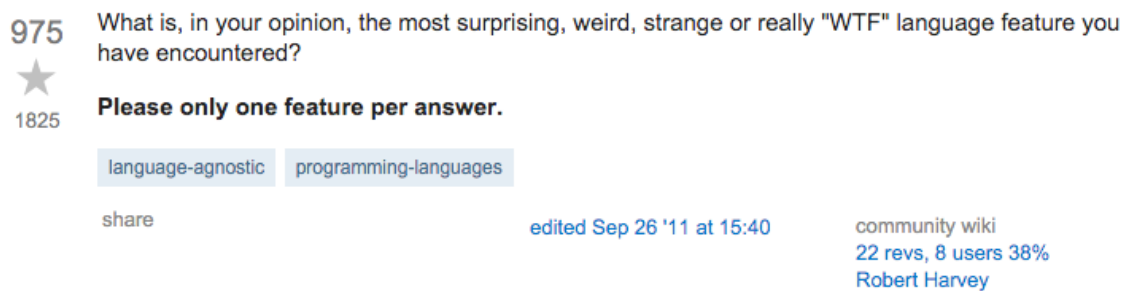


Figure 3.20: Example of StackOverflow question (Harvey, 2011) where people are sharing trivia

Reddit also has more pop culture references in the posts and comments.

All these analyses of StackOverflow and Reddit datasets gives insight into the community structure and user behaviour. This helps to understand technology based PSNs better. They have some similarities and some distinct features. The experiment and analysis helps to answer KQ2.



Figure 3.21: Example of Reddit post(carmichael561, 2013) where people are sharing pop culture references

Chapter 4

The Suman System

The Suman system is designed as a proof of concept to show how Linked Data and Semantic Web technologies can be used to improve search in a Purposive Social Network. The Suman system is built to test the hypothesis of the thesis. It tests if the use of Linked Data and Semantic Web technologies can help with finding answers to unanswered questions in a question and answering systems.

As mentioned in research questions in Chapter 1, the main RQ1 is broken down into smaller questions. The first one is the knowledge question KQ3 that investigates if Semantic Web and Linked Data technologies can be useful in search and discovery of answers in PSNs. The second one is the design problem DP1 that provides solution to RQ1. The prototype system called the Suman system is built to investigate if Linked Data and Semantic Web technologies can be used to answer unanswered questions in PSNs. The main goal of the Suman system is to search for answers to unanswered questions in PSNs. Furthermore, another design problem DP2 is also solved by the Suman system where it recommends experts that could potentially answer the answered questions in PSNs.

This chapter discusses all three research questions, KQ3, DP1 and DP2. It follows the design methodology discussed by (Wieringa, 2014) to do this research. This chapter explains why Semantic Web and Linked Data technologies was used, how the Suman system works, how the model was designed and the search algorithm was built. It is also explained later how an application was built using StackOverflow and Reddit data.

4.1 Use of Semantic Web and Linked Data in Purposive Social Network

This thesis investigates if Semantic Web and Linked Data technologies can help answer unanswered questions in the PSNs. This lead to KQ3 that investigates if Semantic

Web and Linked Data technologies is useful in PSNs. This question can be answered in multiple ways. In this section current literature and examples of PSNs are studied to investigate what are the main challenges encountered to answer RQ1 and how the Semantic Web and Linked Data technologies can provide solutions to those problems.

KQ3 is answered using the empirical cycle in Chapter 5 but in this section it is answered based on related scientific literature and using real world examples of the benefit of using Linked Data and Semantic Web technologies.

4.1.1 Research problem and challenges

KQ3 investigates if Semantic Web and Linked Data technologies are useful to answer unanswered questions in PSNs. This is tested by building the Suman system that takes unanswered questions from PSNs and search for answers and experts that could potentially provide a solution to the unanswered questions.

Firstly, usefulness of Semantic Web and Linked Data in designing the Suman system needs to be studied. Secondly, it needs to be tested that the knowledge added by the use of Semantic Web and Linked Data technology is good.

The first part is discussed in this section and the second part is tested in Chapter 5. To test the first section, the challenges encountered by building the Suman system are analysed. If all the challenges are met, then it validates the use of Semantic Web and Linked Data technologies.

The main design and research challenge to build the Suman system are discussed below.

- **Data Collection and integration:** The study of PSNs requires user data from different online social networks, forums and communities. Some of forums and boards are open to the public and data can be harvested freely using the API and simple screen scraping. It becomes difficult to study PSN when they are closed due to privacy policies and data is not readily available to the general public for consumption.

The focus of this research is Q&A systems and most of the information is crowd-sourced and the knowledge base is created by the contribution of multiple users. The details of each user's contribution are also required to create a proper user-model to find the experts.

- **Name Entity Disambiguation and Linking the Data:** Data from different websites are harvested and integrated together. Integration of data from multiple sources causes name entity problem for main topics. Users might also have multiple accounts in different system, so the information from all these different accounts need to be integrated to form a complete and homogeneous user profile. Same entities

might have different names and data providers might have adopted different URIs for them. This makes it paramount to solve the co-reference and name ambiguity problem (Glaser et al., 2007). Solving the name entity problem is also integral to link the data to the Linked Data Cloud because the topics should be linked with the right concepts in the Linked Data Cloud.

- **Search and query:** The main challenge of the research is to find answers to unanswered questions in PSNs as well as right experts. This is a big problem because first the main topic of the questions needs to be determined and then based on that right answer need to be searched. This also raises the issue of indexing and ranking of search results.
- **Purposive Social Network formation:** The concept of a network is very broad, it can be associated with any kind of relation between people or entities. PSN can be a small, temporary community of people solving a problem and then dispersing once the task is finished. The system should help bring together people with similar interest who can help solve problems.

If the Semantic Web and Linked Data technologies can solve this problem, it would justify the use of these technologies in PSNs to answer unanswered questions.

4.1.2 Benefit of Using Semantic Web and Linked Data in Purposive Social Network

Solving these research challenges will validate the use of Semantic Web and Linked Data technologies. There are systems and algorithms that can search for information without using the Linked Data and Semantic Web technologies. The usefulness of Linked Data and Semantic Web technologies in search and discovery of information has been highlighted in details in Chapter 2 and it is supported by the literature review. Linked Data and Semantic Web help with formatting and structuring the PSN data. Semantic Web technologies eases the sharing and the portability of users' data and the forming of an open, decentralized social network across platforms and websites. The Suman system described in this thesis uses Semantic Web technologies to analyze PSNs. The benefits of using Semantic Web and Linked Data technologies in forming agile PSNs are discussed below.

4.1.2.1 Structured Data

With Linked Data and Semantic Web, every resource is represented by a dereferenceable URI. This URI can be resolved in a document described in RDF format with metadata and ontology to describe its properties and provide definition. It provides the context

of data and people and their interest. It provides a structured knowledge base that is useful in finding correct information and people while forming a community.

4.1.2.2 Linking People to People and People to Data

Using vocabulary like FOAF, people are linked with their friends and colleagues and SIOC profiles link people to their data and users' datasets are linked with each other. This helps to link people to people, people to data and data to data. The interlinked dataset when queried can provide related information.

4.1.2.3 Multidimensional Network and Graph

All data in the Linked Data Cloud are linked with each other based on semantic equivalence and reuse of information by means of URIs. It forms a network graph. This network is formed of people linked with other people by their friendships and other relationships. People are also linked with their data by means of properties. This creates linked network that can be queried across domain and helps to find important information.

4.1.2.4 Integrated Knowledge

Semantic Web technologies can be used to describe different communities and networks. The underlying semantics in the data helps in linking similar concepts. This linking can be integrated together from different datasets to form a large knowledge base.

Each PSN has its own databases. If the data is open and interlinked with other datasets, it can be queried. Relationships can be created between the datasets from different sources and it could be integrated like the Linked Data Cloud.

4.1.2.5 Smart Query and Search

The Linked Data can be queried using SPARQL, or browsed using a semantically enabled browser. Querying the data makes it easy to create many useful applications, mash-ups and discover relationships and patterns. Even if the data or clusters of resources are decentralized, it can be used queried and analysed to measure network structure, network growth and other relationships. Many interesting applications and communities can be generated by analysing the semantic relationship of the data.

4.1.2.6 Social Network Analysis

With the integration of multiple social networks and the elimination of identity fragmentation, social networks and their data can be analyzed and visualized in many ways. (Ereteo et al., 2009) states that classical social network analysis uses only raw social network data and loses some knowledge in the process. Semantic Web frameworks help in mitigating this problem of representing and exchanging knowledge on such social networks with Semantic Web technologies like rich typed graph model (RDF), a query language (SPARQL) and schema definition frameworks (RDFS and OWL). Experiments can be performed to understand the underlying meaning in the network data, how networks are formed, or how the data flows within the network. It can also be used to measure the strength of a relationship or the degree of recognition of experts in networks and much more (Mika, 2005; Jamali and Abolhassani, 2006; Gruber, 2008).

The implementation of the Semantic Web and Linked Data technologies in the PSN is shown in detail in the next section. The Suman system utilises these technologies on PSNs like StackOverflow and Reddit. The usefulness of the knowledge and semantics added in StackOverflow and Reddit is tested and measured by a user experiment and the details can be found in Chapter 5.

4.1.3 Research Validation

The use of Semantic Web and Linked Data technologies will be validated if it meet all the research challenges discussed above. This is a subjective question because these research challenges could be partially met or fully.

The benefit of using Linked data and Semantic Web technologies can also be measured objectively by doing an experiment to test if the added meaning to the detests are useful. This is done in the next chapter by a user experiment. The inference of the experiment gives a strong indication and it could potentially justify use of this technology in PSNs.

4.1.4 Research execution

The PSNs data used in this thesis are from StackOverflow and Reddit. These websites are chosen because these websites are open and have APIs to harvest the data. This meets the first research challenge.

The second research challenge of Name Entitiy Disambiguation is solved by the help of Wikipedia-miner ¹ (Milne and Witten, 2012) and OpenCalais ² (Reuters, 2008) toolkit. These tool use natural language processing to find the main keywords from the text

¹<http://wikipedia-miner.cms.waikato.ac.nz/>

²<http://www.opencalais.com/>

and use machine learning algorithms to match the keywords to topics and particular vocabulary. The main entities are matched with Wikipedia articles and OpenCalais vocabulary. These entities are linked with DBpedia and OpenCalais knowledge base. The keywords are also linked with DBpedia categories to add more relationships and semantics.

The third research challenge of searching and querying answers for unanswered questions is done by creating the Suman system that utilises the semantically enriched data. This is done by converting all the PSNs data into RDF and storing it into the database. The recognized keywords are also given a confidence score and stored into a document keyword graph. All of this uses Semantic Web technologies to perform search and query.

The final research challenge is to build and study PSNs. The relation between people and objects are varied with different attributes. When users do not create an explicit relationship, the relationships are created by studying their communication networks (posts, questions, answers, etc.). This also creates a network of experts based on topics and categories. Attributes like relationship/links between users, and users and objects that connect people and entities are analyzed. The incentive model of the website that encourages people to collaborate and contribute is also studied. The structure of the communication and network growth over time is also analyzed (Mika, 2005). Studying this structure is made easier because of linking of data with experts and categories.

So, all in all the use of Semantic Web and Linked data facilitates building of the Suman system.

4.1.5 Result evaluation

Chapter 5 discusses in details the user experiment to test the usefulness of adding semantics to PSN data. In this section the use of these technologies is justified to show why they are used and what is the best possible way to employ Semantic Web and Linked Data.

The data from StackOverflow and Reddit are from the technology domain and have structured knowledge base. The existing ontology and knowledgebase on open Linked Data consists of many data sources from technology domain. Adding semantics in StackOverflow and Reddit data source makes it machine readable and adds knowledge. This helps to identify the meaning and core topics in the datasets. Furthermore, Linked Data links this topic to the other datasets, thus adding more information and helping to categorise and organize the datasets.

So, using Semantic Web and Linked Data technologies is a useful idea in the PSNs. This helps to understand the underlying meaning and concepts of the question and answers. This added semantic could potentially improve the search and discovery of information.

Then the question arises, what is the best way to employ Semantic Web and Linked Data to the PSNs datasets. This could be done by doing a Name entity disambiguation to the questions and answers to understand the main topics and concepts of the question and answers. This adds semantics to the datasets. These topics could be linked to the Linked Data Cloud to help form connections between multiple datasets and use the relationship to infer more information. The added links provide additional knowledge and information that could potentially help to improve search and discovery of information.

The usefulness of the semantics and linking needs to be measured to justify using it. This is done using a user experiment where participants look at the added semantics and rate how well these keywords describe the question and answers. This is discussed in detail in Chapter 5.

The other way to measure the usefulness of the added semantics is to analyse the relationship and categories of the topics in the questions and answers. These relationships and categories help to categorize and structure the knowledge and help to find similar questions and answers. It is hypothesized that this helps in improving search and discovery of answers to the unanswered questions. The usefulness of this could be measured by another user experiment where participants can rate how well the search result can answer the unanswered question. Statistical analysis is done on the data collected from these experiments. This is discussed in detail in Chapter 5.

If the finding and results are statistically significant, then it could be inferred that using Semantic Web and Linked Data improve the PSNs and answer the research questions. This would justify the use of these technologies and we could be satisfied with the end result because it meets the research goals and answers KQ3.

4.2 The Suman System

The Suman system is built as a prototype to answer the main research question. It is designed as a proof of concept to show how Linked Data and Semantic Web technologies can be used to improve search in a Purposive Social Network. It tests if the use of Linked Data and Semantic Web technologies can help with finding answers to unanswered questions in a question and answering systems. The Suman system also implements a semantic query algorithm that uses the semantics and crowdsourced community information together to improve search and recommends users. This could potentially solve DP1 and DP2.

This section explains how the Suman system works, how the model was designed and the search algorithm was built. It is also explained later how an application was built

using StackOverflow and Reddit data. This is explained by using design science research methodologies. This section is structured using the design science methodologies engineering cycle.

4.2.1 Problem Investigation

This thesis aims to solve the problem of unanswered questions in PSNs using Linked Data and Semantic Web technologies. In the earlier section use of Linked Data and Semantic Web technologies PSN is justified.

In this section the two design problems are discussed. The first design problem (DP1) is to find answers to unanswered questions in PSNs. This means that a prototype system should be built that takes questions, answers, user profiles and other crowdsourced data (votes, favourites, tags, etc.) from PSNs and use these data to provide answers to unanswered questions. The second design problem (DP2) this thesis aims to solve is to find experts in the field that could potentially answer the unanswered questions in PSNs.

A solution based approach is required to solve this problem and the Suman system is designed to provide a solution to this problem.

The main stakeholders for this research problems are the users of PSNs that ask questions and do not get any answers.

The main artifacts to solve this problem is the Suman system. It's a prototype system that uses Semantic Web and Linked Data technologies to search for answers for the unanswered questions and recommend experts that could potentially answer those questions.

The main goal and requirement is to find answers to the unanswered questions in PSNs using Linked Data and Semantic Web technologies and a recommended list of experts in the field. This would answer both DP1 and DP2.

4.2.2 Treatment Design of the Suman System

The main requirement to solve the research question is to build a system that could find answers to the unanswered questions using Linked Data and Semantic Web technologies and recommend experts. This could be done in the following ways:

1. Generating the answers to the questions by the Suman system automatically. If the system meets this goal, it will fully meet the design goals.

2. Searching for answers to the questions from the pre-existing knowledge base. This would require collecting the pre-existing data in the PSNs and solving the design challenges discussed in Chapter 1 (research challenges section). If the system meets this goal, it will fully meet the design goals.
3. Searching for experts in the PSNs that could answer the unanswered questions. This would not directly provide an answer to the questions, but it'll find the right experts who could potentially answer the questions. This does not meet the research question fully, only partially and it's subjected to failures.

For the Suman system, the second and third approach is taken to solve the design problem. The Suman system collects PSNs data and uses it to search for answers that could provide a solution to the unanswered questions. If there are no good answers to the unanswered question experts are recommended that could potentially answer the question. The Suman system solves all the research challenges discussed in chapter 1 section 1.3.2.1. It is discussed in details in chapter 4 section 4.1.1.

So, the main goals of the research and the Suman system are, firstly, to search for answers to unanswered questions in PSNs. Secondly, to search for experts that could possibly answer the unanswered questions in the PSNs.

To solve these issues, the main components of the Suman system are divided into five parts- data collection, data structuring, annotation and linking, database and query, and search and expert recommendation. Figure 4.1 gives an overall view of different components and the system design.

4.2.2.1 Data Collection

The Suman system collects PSNs data to build its knowledge base. The datasets usually consist of questions, answers, user details and other crowdsourced data (votes, favourites, etc.) from Q&A systems. The data could be collected from APIs, screen scraping or data dumps. The data can be collected from one or many PSNs.

There are two main types of data collected in the Suman system. Firstly, the post data, this includes questions, answers, tags, categories with votes and comments. This helps to create the knowledge base of questions and answers that can be used for search. Secondly, users' profile and activity data are collected. This includes users' details (name, username, email, etc.) and their activity (posts, votes, favourites, badges, reputation). This helps to create a user model and that leads to recommend experts.

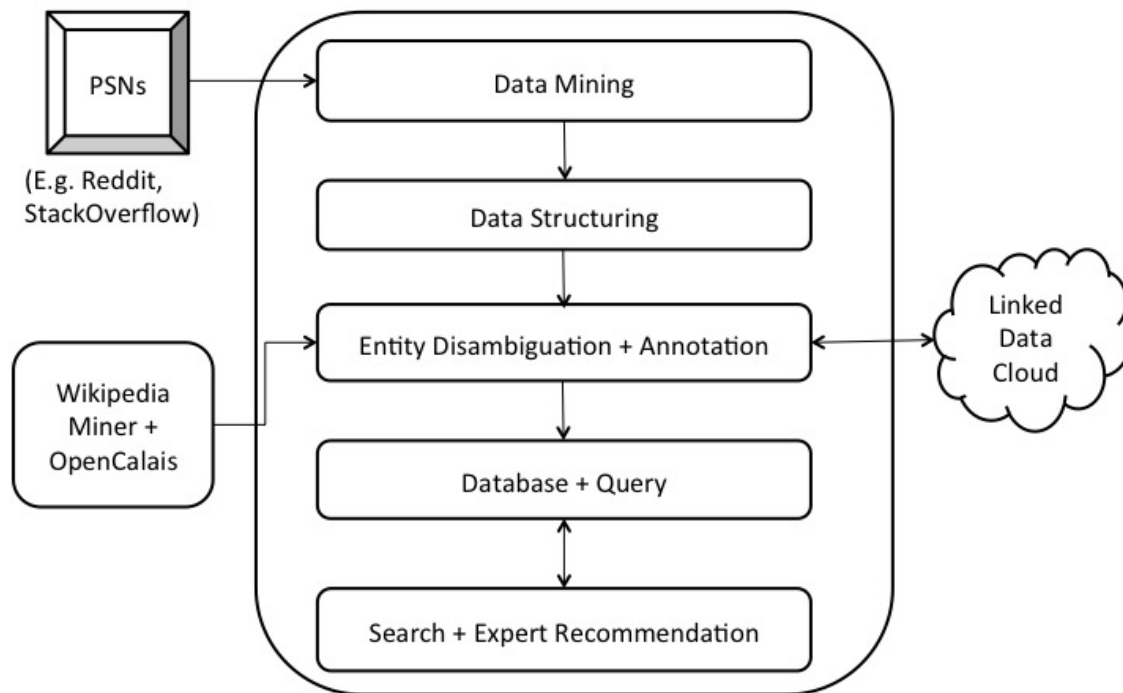


Figure 4.1: The Suman system design components.

4.2.2.2 Data Structuring

The data collected from different PSNs might be structured differently. They might have different fields and different data types. The second part of the Suman system takes heterogeneous data and gives it a common structure using Semantic Web technologies.

The Semantic Web requires the data to be structured properly using some schema to give data meaning. For this system, the data were turned into RDF. To describe the data in RDF, an ontology is used to define concepts and relationships. FOAF ontology is used to describe the users and their profile information. Similarly, the posts (questions, answers, comments, etc.) are described using SIOC and DC ontology. Any properties that can't be describes using these vocabularies are described by creating own schema in RDFS.

4.2.2.3 Keyword Annotation and Linking

One of the main benefits of using Semantic Web and Linked Data technologies are adding semantics to the data and linking it to other datasets. Adding semantics helps by providing the context and domain specific meaning to the data. Linking the data helps to find related information about a resource.

The main challenge of adding semantics to the data is name entity disambiguation as discussed in Chapter 2. Once names and entities are resolved, the context of the domain of the entities could be used to add extra semantics, categories and other relevant information to the resources. For example, 'Eclipse' could mean a natural phenomenon or an IDE (Integrated Development Environment). Knowing the domain of the data, if it's from technology domain or general domain, allows us to add proper categories and semantics to the entities.

The third step of the Suman system adds keywords to the collected datasets from technology domain. It also adds categories to the keywords to link them together and form semantic relationships between them. These data with added categories and keywords are later linked to DBpedia and OpenCalais datasets to find related information. DBpedia is the structured data from Wikipedia. It allows other datasets to link to the Wikipedia data.

DBpedia knowledge base describes 4.58 million things (Auer et al., 2007). OpenCalais is Thompson Reuters initiative that tags keywords, topics, etc (Reuters, 2008). This is done by using two tools, Wikipedia Miner and OpenCalais. These tools do the name entity recognition and match entities with topics and categories. Wikipedia Miner and OpenCalais do not convert the resulted keywords into Linked Data or RDF. The returned keywords were further structured into RDF and linked with DBpedia and OpenCalais datasets. The linking of datasets helps to find related information about the topics.

Wikipedia Miner and OpenCalais tools are used to do the name entity recognition and match it to a known vocabulary and taxonomy. (Li et al., 2003; Glaser et al., 2009) states that adding semantics from different data sources improves the quality of the metadata and semantic context significantly. It also overcomes any false match made by one application. Hence, both OpenCalais and DBpedia dataset were used to resolve the name entity issue and link the data. Wikipedia Miner and OpenCalais do natural language processing of the text and annotate with keywords. The annotations are then matched to the Wikipedia topics and OpenCalais entities.

The Wikipedia Miner service uses word sense for disambiguation. The algorithm uses machine learning to detect key terms in a text and disambiguate them against Wikipedia articles. Wikipedia Miner provides a JAVA API to access the Wikipedia database, including all categories (Milne and Witten, 2012).

The Wikipedia categories are attached to the keywords. Every keyword is linked to its parent and child categories. This will later help in expanding keywords with other related keywords and other keywords belonging in the same category. This results in a graph. A keyword can be part of multiple categories, so a keyword can have multiple parent categories.

OpenCalais tool creates a semantically tagged metadata for the content using natural language processing and machine learning algorithm. It provides many features like tag integration with different taxonomy and vocabulary, Geo-mapping of location and semantic annotation of keywords (Corlosquet et al., 2009). The keywords extracted by OpenCalais is also structured into RDF and linked to the OpenCalais dataset.

4.2.2.4 Database and Query

In the previous steps all the posts, users' data and added keywords from Wikipedia Miner and OpenCalais are structured into RDF. This RDF is stored in the Stardog database and it is also referred as documents. Stardog is a semantic graph database. It supports RDF graph data model and SPARQL query language. It supports OWL 2 and user defined rules for inference, reasoning and constraints. It uses HTTP protocol and provides with SPARQL endpoint for applications to perform queries.

Stardog (Inc, 2015) has a R2RML mapping feature to create virtual graphs. This maps the Stardog graph to external data source so it could also query the external data source. It has many useful features. It is easy to import and export data. Stardog provides a web console for easy browsing. The web console provides admin controls for managing the database. It also provides options to browse and search the graph on the web console.

Stardog includes an RDF-aware semantic search function. It indexes RDF literals and supports information retrieval style queries. The indexing strategy creates a search document per RDF literal. It also builds an extra index for named graph which are used when SPARQL queries specify datasets using 'FROM' and 'FROM NAMED'. The indexing strategy can be changed by the admin and indexing parameters are configurable (Inc, 2015).

Stardog uses Lucene's default text analyzer for searching information. It also provides a custom analyzer to customize the Lucene analyzer to support different natural languages, indexing, domain specific word lists, etc. Full-text search support is disabled by default, but it can be enabled any time to rebuild the search index and it is also customisable like the RDF index (Inc, 2015).

Stardog supports SPARQL for querying RDF graphs as well as providing functionalities such as searching, obfuscating and browsing graph data. Stardog from version 3.0 and onwards supports federated queries for the 'SERVICE' keyword which allows distributed RDF data sources to be queried. Lucene syntax is allowed to perform search queries. 'textMatch' predicate is used to access the search index and a threshold score can be provided to control the quality of the search results. Stardog scores the results between 0 and 10. The search and query allow Lucene search modifiers (Inc, 2015).

Stardog allows admin to configure many different functions such as search, querying, indexing, scoring, etc. It provides support for browsing and searching the graph on the web console. It provides an API and SPARQL endpoints for accessing and querying its own database. All these benefits helped in choosing Stardog for the Suman system.

4.2.2.4.1 Database Indexing and Configuration: To improve the search and discovery of answers and experts in the Suman system the database index was to optimized. This would theoretically improve the search results.

All the questions, answers and comments have keywords associated with them. These keywords also have categories. Keywords and categories are linked together and create a graph. Every post is associated with keywords and through it to categories.

The Stardog indexing system is configurable. It has RDF aware semantic search functions. It indexes RDF literals and creates a search document per RDF literal. Each document consists of the following fields: literal ID; literal value; and contexts. Stardog uses the Lucene text analyzer to index the database and this analyzer is customizable. It is possible to implement a custom analyzer that Stardog will use by implementing `org.apache.lucene.analysis.Analyzer`. It allows Stardog to support different domain specific knowledge.

The database is customized by adding the keyword-categories graph to the ‘common-gram’ analyzer. This constructs n-grams for frequently occurring keywords. The n-gram is a contiguous sequence of n keywords for each post. This also extends to the categories. The KeywordAnalyzer ‘tokenizes’ the entire stream as a single token. The ‘ngram-analyzer’ also constructs n-gram tokenizers and filters for multiple keywords. One useful feature of the Lucene analyzer is its support of Wikipedia syntax. This helps in using Wikipedia Categories and DBpedia linked set while reasoning.

The Lucene analyzer indexes documents for text based search. It follows the same principle as Latent Semantic Analysis and TF-IDF (Term Frequency- Inverse Document Frequency) Weighting (Salton et al., 1975). There is a bidirectional relationship between documents (questions, answer, comments) and keywords. The frequency of occurrence of each keyword in a document is calculated and the total frequency of documents for each term is also calculated. Since, there are a finite number of keywords in the database, so it is easier to calculate the document frequency for each keyword.

The frequency of keywords in each document is counted and a term frequency matrix is created. In addition, the frequency of all the documents for each keyword is counted and document frequency matrix is created. This is useful because it helps to determine the importance of a keyword in each document which minimises the query run time when searching for documents for particular keywords or set of keywords.

Stardog can perform both SPARQL query search and full text search. It has reasoning capabilities using OWL 2 Direct Semantic Entailment Regime. This means that the semantic conditions are defined with respect to ontology structures. It follows the instances of ontology classes as defined in the OWL 2 syntax specification (Glimm, 2012). It performs reasoning in lazy and late-binding fashion. The reasoning is performed at query time according to user-specified reasoning type. It does not materialize inferences (Inc, 2015). The reasoning types could be enabled as it is disabled by default.

In order to perform query evaluation with reasoning, Stardog requires an RDF Schema to be present in the database. The RDFS file is imported along with the RDF files containing all the data. Stardog supports an in-built query rewriting via predefined rules. It later executes the resulting expanded query against the data normally.

4.2.2.5 Suman Search Algorithm

Stardog provides a SPARQL endpoint for the application to use. The Suman search algorithm uses the SPARQL endpoint and queries it over HTTP client.

The Suman system takes the unanswered questions from Q&A systems and finds similar questions with answers. This was done to test the hypothesis to see if the resulted answers could be used to answer the original unanswered questions. The Suman system also recommends experts who could answer the questions.

The Suman search algorithm is discussed in detail below. The notations used are as follows:

1. K = set of keywords
2. D = set of documents
3. Q = Query
4. V = Vote
5. S = Score $S \in \mathbb{R} : 0 \leq S \leq 10$

Here, questions and answers are referred as documents. Each document has a set of keywords K associated with it. The system searches for answers to an unanswered question. Hence, a query Q consists of an unanswered question and the keywords associated with it. Vote V referred to the vote given to questions and answers by the users. Score S is the score given to the search result based on its validity.

Algorithm 1 The Suman Search Algorithm

```

1:  $D = [], K = [], minScore = 0.7$ 
2:  $[\bar{q}, K] \leftarrow FindQuestion(Random)$ 
3:  $Q = k_1 \wedge k_2 \dots \wedge k_n \forall k_i \in K$ 
4:  $[D, S] \leftarrow FindDocs(Q)$ 
5:  $EK \leftarrow Expand(K)$ 
6:  $\bar{Q} \leftarrow ek_1 \vee ek_2 \dots \vee ek_n \forall ek_i \in EK$ 
7:  $[ED, ES] \leftarrow UpdateDoc(D, \bar{Q}, minScore)$ 
8:  $[ED, S^{\bar{Q}}] \leftarrow TextSearch(\bar{q}, ED)$ 
9:  $[ED, S] \leftarrow UpdateScore(ED, ES, S^{\bar{Q}})$ 
10:  $[ED, V] \leftarrow GetVote(ED)$ 
11:  $[D^{Final}] \leftarrow ScaleScore(ED, S, V)$ 
12:  $[D^{*Final}] \leftarrow GetContext(D)$ 

```

```

1: procedure EXPAND(K)
2:    $EK \leftarrow K$ 
3:   for  $\forall k \in K$  do
4:     if  $k$  has parent  $p$  then
5:        $EK \leftarrow EK \cup p$ 
6:     else  $k$  has children  $C$ 
7:        $EK \leftarrow EK \cup C$ 
8:     end if
9:   end for
10: end procedure

```

4.2.2.5.1 Detailed Explanation of Each Step

1. $D=[], K = [], minScore = 0.7$ = D is an empty list and stores all the documents returned after running the query. This is referred as cache in this section because the database stores it in the cache memory. K is an empty list that stores all the keywords associated with the document. The minScore variable stores the value of threshold score. In this algorithm, it is 0.7. Any documents with scores lower than the minScore is pruned to maintain the quality of the search result.
2. FindQuestion(Random) = This operation returns a randomly selected unanswered question \bar{q} from the database. This question has a set of keywords K associated with it. Those keywords are also retrieved for the query.
3. Query(Q) = The first query is created by using all the keywords associated with the unanswered question. The keywords are joined using the \wedge (AND) operator to make sure the search result consists of all the keywords.
4. FindDocs(KeywordSet) = This operation takes the sets of keywords K related to the query question Q. Then it uses the set of keywords to run a SPARQL query and search for documents that contain all the keywords. The SPARQL query is created automatically by the Suman system. This operation creates a cache file [D,

S] that holds all the returned results with a score associated with each result. The query gives a higher score to documents that have all the keywords and multiple occurrence of the keywords associated to it. Any documents with a score less than minScore are pruned.

5. $\text{Expand}(\text{KeywordSet})$ = This operation takes a set of keywords K associated with the query question Q and returns an *expanded* set of keywords as described in the next algorithm. It takes each keyword and finds the parent keyword if it has one. As described in section 4.1.3 each keyword has been disambiguated earlier and has a broader and narrower term associated with it. The broader terms are the parent keyword and the narrower terms are the children keywords. If the keyword has a parent keyword associated with it, then that keyword is added to the keyword set. If there are no parents, then the children term is searched and added to the keyword set if they exist. This generates an expanded set of keywords EK .
6. $\text{UpdatedQuery}(\bar{Q})$ = The new query \bar{Q} is updated by adding all the expanded set of keywords joined using the \vee (OR) operator. This is done to get documents that have most of the keywords associated with it but not all.
7. $\text{UpdateDoc}(\text{DocumentSet}, \text{UpdatedQuery}, \text{minScore})$ = This operation takes a set of documents D , a query that consists of the set of expanded keywords \bar{Q} , and a score called minScore (optional). If the minScore is greater than 0.7 and the documents already exist in the cache file, then the document's score are updated to give it a boost. Otherwise the documents are added to the cache file with its score.
8. $\text{TextSearch}(\text{Query}, \text{DocSet})$ = This operation takes the text of the questions \bar{q} and performs a text search using the Lucene search engine in the cache dataset. This search is not done on the whole database but the cached dataset that consist of all the documents related to the keywords. This operation returns a new score for the documents contained in the cache.
9. $\text{UpdateScore}(\text{DocumentSet}, \text{KeywordScore}, \text{TextSearchScore})$ = The cached documents have two scores. The first score was given after the keyword query results and the second score was given after the text search results. This operation updates the score by taking the average of the both scores. This is done to normalize the score and keep it within the range of 0.7 to 10.
10. $\text{GetVote}(\text{DocumentSet})$ = This operation gets the votes received to each document in the cache.
11. $\text{ScaleScore}(\text{DocumentSet}, \text{Score}, \text{Votes})$ = This operation takes the votes V of the documents and scale it in the same range as score S . The scores are usually between 0.7 and 10 and votes are \mathbb{Z} . The votes are normalized to be $V^* \in \mathbb{R} : 0 \leq V^* \leq 10$. This is done by unity based normalization (von Davier, 2010) and the value is

restricted in the range between a and b . Here, the range of a and b are same as the range of score S .

$$\bar{V} = a + \frac{(V - V_{min})(b - a)}{V_{max} - V_{min}}$$

For example, if the cache had three documents with votes 10, 15 and 20 then unity based normalization turns their score into 0.7, 5.35 and 10 respectively to keep the score within the range of 0.7 and 10. The Suman algorithm does this with all the documents present in the cache.

If the votes were negative, then the normalized vote value is deducted from score and if they were positive, normalized vote value is added to the score. Then the average is taken of the score and the normalized value of votes. Any documents with final score less than 0.7 are pruned from the list. This modifies the final ranking of the documents and the new sorted list of the documents is presented as the final result.

12. `GetContext(DocumentSet)` = This operation retrieves the parent posts of each document, if available. The questions do not have parent posts, but answers and comments have parent posts associated with them. All the parent posts of answers and comments are retrieved to provide context to the final result.

The final result consists of a ranked list of documents that could potentially answer the unanswered questions. The top ten results are shown as a query result. Number of results could be modified to show more or all of the results. The results consist of answers with the parent posts.

This search algorithm used keyword based semantic search and combined it with the text based search. The final ranking of the search result was modified using the crowdsourced votes given for questions and answers by the community.

The first step of searching the answers is using the keyword graph to find the best match. The query result is limited to the score greater than 0.7. If there are results with score more than 0.7 then the votes of the answers are used to sort them. Top ten results of the returned query with the original question is shown as the final result of the query.

If in the initial query there are no results with the score greater than 0.7 then the search term is expanded from keywords to include the categories of the keywords. This helps to find the similar questions and answers in the same programming categories, they might not be about the exact topic. Again, the votes of each answer are used to sort them and top ten results are shown. If there are no answers with score more than 0.7 then no results are shown in the application.

4.2.2.6 Expert Finder

The questions, answers and keyword graph are extended to users. Users are linked with the keywords and votes to create a user keyword graph. Every document (questions, answers, posts, comments) has a user associated with it. This helps in joining the keywords with the users. Also, keywords have categories related to them, it makes a graph of users related to categories too. This helps in recommending experts.

The Suman expert recommendation algorithm is discussed in detail below. The notations used are as follows:

1. K = set of keywords
2. E = set of experts
3. Q = Query
4. R = Reputation points of experts
5. S = Score $S \in \mathbb{R} : 0 \leq S \leq 10$

Algorithm 2 The Suman Expert Recommendation Algorithm

```

1:  $E = [], K = [], minScore = 0.7$ 
2:  $[\bar{q}, K] \leftarrow FindQuestion(Random)$ 
3:  $Q = k_1 \wedge k_2 \dots \wedge k_n \forall k_i \in K$ 
4:  $[E, S] \leftarrow FindExperts(Q)$ 
5:  $EK \leftarrow Expand(K)$ 
6:  $\bar{Q} \leftarrow ek_1 \vee ek_2 \dots \vee ek_n \forall ek_i \in EK$ 
7:  $[EE, ES] \leftarrow UpdateExperts(E, \bar{Q}, minScore)$ 
8:  $[EE, R] \leftarrow GetReputation(EE)$ 
9:  $[E^{Final}] \leftarrow ScaleScore(EE, S, R)$ 
10:  $[D^{*Final}] \leftarrow GetDetail(E)$ 

```

```

1: procedure EXPAND( $K$ )
2:    $EK \leftarrow K$ 
3:   for  $\forall k \in K$  do
4:     if  $k$  has parent  $p$  then
5:        $EK \leftarrow EK \cup p$ 
6:     else  $k$  has children  $C$ 
7:        $EK \leftarrow EK \cup C$ 
8:     end if
9:   end for
10: end procedure

```

4.2.2.6.1 Detailed Explanation of Each Step

1. $E=[], K=[], \text{minScore} = 0.7$ = E is an empty list and stores all the experts returned after running the query. K is an empty list that stores all the keywords associated with the document. The minScore variable stores the value of threshold score. In this algorithm, it is 0.7. Any documents with scores lower than the minScore is pruned to maintain the quality of the search result.
2. $\text{FindQuestion}(\text{Random})$ = This operation returns a randomly selected unanswered question \bar{q} from the database. This question has a set of keywords K associated with it. Those keywords are also retrieved for the query. This question is same as the question used for search algorithm.
3. $\text{Query}(Q)$ = The first query is created by using all the keywords associated with the unanswered question. The keywords are joined using the \wedge (AND) operator to make sure the search result consists of all the keywords.
4. $\text{FindExperts}(\text{KeywordSet})$ = This operation takes the sets of keywords K related to the query question Q . Then it uses the set of keywords to run a SPARQL query and search for experts that are linked to all the keywords. The SPARQL query is created automatically by the Suman system. This operation creates a file $[E, S]$ that holds all the returned results with a score associated with each result. The query gives a higher score to experts that have all the keywords and multiple occurrence of the keywords associated to them. Any experts with a score less than minScore are pruned.
5. $\text{Expand}(\text{KeywordSet})$ = This operation takes a set of keywords K associated with the query question Q , and returns an *expanded* set of keywords as described in the next algorithm. It takes each keyword and finds the parent keyword if it has one. As described in section 4.1.3 each keyword has been disambiguated earlier and has a broader and narrower term associated with it. The broader terms are the parent keyword and the narrower terms are the children keywords. If the keyword has a parent keyword associated with it, then that keyword is added to the keyword set. If there are no parents then children terms are searched and added to the keyword set if they exist. This generates an expanded set of keywords EK .
6. $\text{UpdatedQuery}(\bar{Q})$ = The new query \bar{Q} is updated by adding all the expanded set of keywords joined using the \vee (OR) operator. This is done to get experts that have most of the keywords associated with them but not all.
7. $\text{UpdateExperts}(\text{ExpertSet}, \text{UpdatedQuery}, \text{minScore})$ = This operation takes a set of experts E , a query that consists of the set of expanded keywords \bar{Q} , and a score called minScore (optional). If the minScore is greater than 0.7 and the expert already exists in the list, then the experts' score are updated to give it a boost. Otherwise the experts are added to the list with their score.

8. $\text{GetReputation}(\text{ExpertSet})$ = This operation gets the reputation points of all the experts on the list.
9. $\text{ScaleScore}(\text{ExpertSet}, \text{Score}, \text{Reputation})$ = This operation takes the reputation R of the experts and scale it in the same range as score S . The scores are usually between 0.7 and 10 and reputations are \mathbb{Z} . The reputation is normalized to be $R \in \mathbb{R} : 0 \leq R \leq 10$. This is done by unity based normalization (von Davier, 2010) and the value is restricted in the range between a and b . Here, the range of a and b are same as the range of score S .

$$\bar{R} = a + \frac{(R - R_{\min})(b - a)}{R_{\max} - R_{\min}}$$

For example, if the cache had three experts with reputations 1000, 1500 and 2000 then unity based normalization turns their score into 0.7, 5.35 and 10 respectively to keep the score within the range of 0.7 and 10. The Suman expert recommendation algorithm does this with all the experts present in the list.

The average is taken of the score and the normalized value of reputations. Any expert with final score less than 0.7 are pruned from the list. This modifies the final ranking of the experts and the new sorted list of the experts is presented as the final result.

10. $\text{GetDetail}(\text{ExpertSet})$ = This operation retrieves all the available details of the experts. This includes their name, username, profile link, email, etc. All these details provide more information about the experts and how to contact them. This provides some context to the final result.

The final result consists of a ranked list of experts that could potentially answer the unanswered questions. The top ten results are shown as a query result. Search of experts is done similarly to the answers. The list of experts is best matched with the keywords of the question so they are assumed to be experts on those topics.

The experts have additional information associated with them like their location, latest activity, posting history, etc. This can be used to modify the query result to find the experts in the same time zone, or experts who were recently active and post on the website. This would potentially help in finding the right experts for the right time and help users to create an agile PSNs. For example, using the response time of the experts will help to recommend experts who quickly respond to questions. Users who are in urgent need for a solution, they can find quick responders instead of the experts with highest reputation points. These features have not been implemented to the Suman system yet, but they can be added in the future that would potentially help to create PSNs quickly and based on different criteria.

4.2.2.7 Design Innovation in the Suman System

The Suman System uses PSNs data, specifically question and answering systems like StackOverflow and Reddit. These PSNs are online communities where people post questions and other people answers those questions. Sometimes, users also share other information like links, blog posts, etc.

Some of the PSNs like StackOverflow are not a social network as people cannot create friendships or other ties but in some cases like Reddit users can create friendships and follow their friend's work.

The Suman system does not provide a platform for creating friendship or other ties as its primary objective is search and information retrieval. But it does use users' communication structure and interests to determine how closely they are related to each other based on topics of interest. This is not a platform to create a social network, but the Suman system utilizes the community structure and the user's expertise and interests to recommend experts that could potentially solve the problem and answer the unanswered questions. This feature of the Suman system makes it purposive, because it helps to solve the problem and it makes it social because it recommends users who you can connect with and talk to get solution for your problems.

Furthermore, the Suman system finds the main topics for questions and answers and recommends users based on those topics, thus recommending experts with very focused interest. These features of the Suman system make it a system that could potentially be used to create an agile PSN where recommended experts could form communities and help solve particular problems.

4.2.2.7.1 Special feature of the Suman algorithms: The Suman system is a new contribution to the research community as it is trying to solve a specific problem. It uses PSNs data and helps to find answers to unanswered questions and recommend experts who could potentially answer the questions. The Suman search algorithm is special in this specific reason, because it uses the PSNs crowdsourced datasets. This includes questions, answers, user profile, votes, favourites, etc. to make a weighted keyword graph. This algorithm also uses a combination of keyword based text search and keyword based semantic search. This approach helps to expand the categories of the question topics and finds answers from related topics that could potentially solve the problem.

The Suman system also created a weighted document keyword graph to improve search. The graph linked the keywords with all the questions, answers and experts. The links were weighted by the votes given to questions, answers and the frequency of the keywords. The data were used to improve the indexing of the dataset and rank the SPARQL search results.

The other special feature of the Suman system is the way it ranks the search results. It uses the ranking of the search result based on semantic analysis, and it also uses the crowdsourced data like votes and favourites to rank the search results. This combination of search and ranking over a specific dataset helps to find the answers and experts makes the Suman search algorithm different and special from other search algorithms.

4.2.3 Design Validation

The Suman system has specific requirements to search for answers to unanswered questions from PSNs and recommends experts who could potentially answer the questions. The Suman system uses Linked Data and Semantic Web technologies to do this. The system will be considered valid if it meets the above mentioned requirements.

The Suman system needs to be validated to check the it satisfies the design criteria and meets the research goals.

4.2.3.1 Validating the Suman system design

The Suman system searches for answers to the questions from the preexisting knowledge base in PSNs. To successfully use the semantic search approach the Suman system needs to meet all the research challenges discussed in Chapter 1.

This is a valid design approach because the Suman system is a real world prototype that uses real world data from PSNs, specifically StackOverflow and Reddit. It uses multiple websites to collect and harvest the data and the system will use real world unanswered questions and search for answers. This uses technical action research method to answer the research questions.

Furthermore, the Suman system uses available research tools to solve the Name Entity Disambiguation problem and the search algorithm is designed to specifically solve the problem discussed in the research question. Solving these research challenges will validate the design treatment. The Suman system if answers the research questions could meet all the design requirement by solving the research problems mentioned above.

There are certain trade-offs when choosing to solve the design problem using existing tools and Semantic Web technologies, but the system meets the final goal of finding answers to the unanswered questions. Furthermore, the outputs of the Suman system are tested in Chapter 5 and answers KQ3, KQ4 and KQ5.

The system could also be evaluated using the existing unanswered questions to see if it searches for answers and experts. If this evaluation is successful, the Suman system design is considered valid and it fully answers the research question and meets the DP1 and DP2 goals.

4.2.4 Treatment Implementation using StackOverflow and Reddit Datasets

The Suman system can work with any PSNs that are questions and answering systems. For this thesis, StackOverflow and Reddit websites were chosen to make an application using the Suman System. These websites are good examples of PSN and they have APIs to get their data. The data are public and open for the users to extract and analyze. Both websites have a focused interest, active communities and use crowdsourcing to solve technology related problems.

StackOverflow is a question and answering website in which programmers ask questions and other programmers who have some expertise in the field answer them. This is a popular website, ranked 57 globally in Alexa ranking(Alexa, 2015b) as of November 2015. StackOverflow has lots of users, questions and answers. Reddit is another popular website ranked 32 globally in Alexa ranking(Alexa, 2015a) as of November 2015. In the Reddit people share news, information and have discussions. The Reddit community is divided and organized based on interest. These divisions are called subreddits. Data from programming based subreddits were collected for the Suman system.

4.2.4.1 StackOverflow Data Mining

The StackOverflow's data are public and has an API to retrieve data. StackOverflow also provides a regular data dump of all their public data which is used in creative common license³. The data dump consists of some public data like questions, answers and some information in users' profiles and other information are anonymized. The votes and favourites data are anonymous. These votes and favourites give insight on the quality of questions and answers, but an important aspect of linking users to their votes is lost. Users' voting information and favourites post information could have given some insight on their behaviour and helped us to create a link between users with similar behaviour. The information could also have been used to link users' votes with other users' responses, hence extending the communication network of the community.

The first method used to get the majority of the data is the data dump. StackOverflow releases a new data dump quarterly. The dump has data about questions, answers, comments, user information (public data only), badges and votes. The data are in XML format and the files are zipped and shared using P2P torrent. A data dump until June 2014 was downloaded for this experiment.

The biggest problem faced while processing the XML files was the size. The whole data dump was 35 GB and the biggest file was the posts XML file sized 21 GB. To overcome this problem the files were broken down into smaller files to parse the XML and store it

³<http://blog.stackoverflow.com/2009/06/stack-overflow-creative-commons-data-dump/>

in a database. The other issue with the dataset was the encoding. The parser encoded all the text to Unicode and then stored it in the Suman database.

The API was used to get lists of tags, total number of questions in each tags, tag synonyms and related tags. The API returns JSON files; these files were parsed and stored in the database. The API also had a limit of 30 calls per minute and the IP address without an access token can make a total of 10,000 calls. That is why the data dump was used to get core data and the API was used to get specialized data that was not available in the data dump.

Figure 4.2 shows StackOverflow dataset schema. This was built to store the data temporarily in MySQL database.

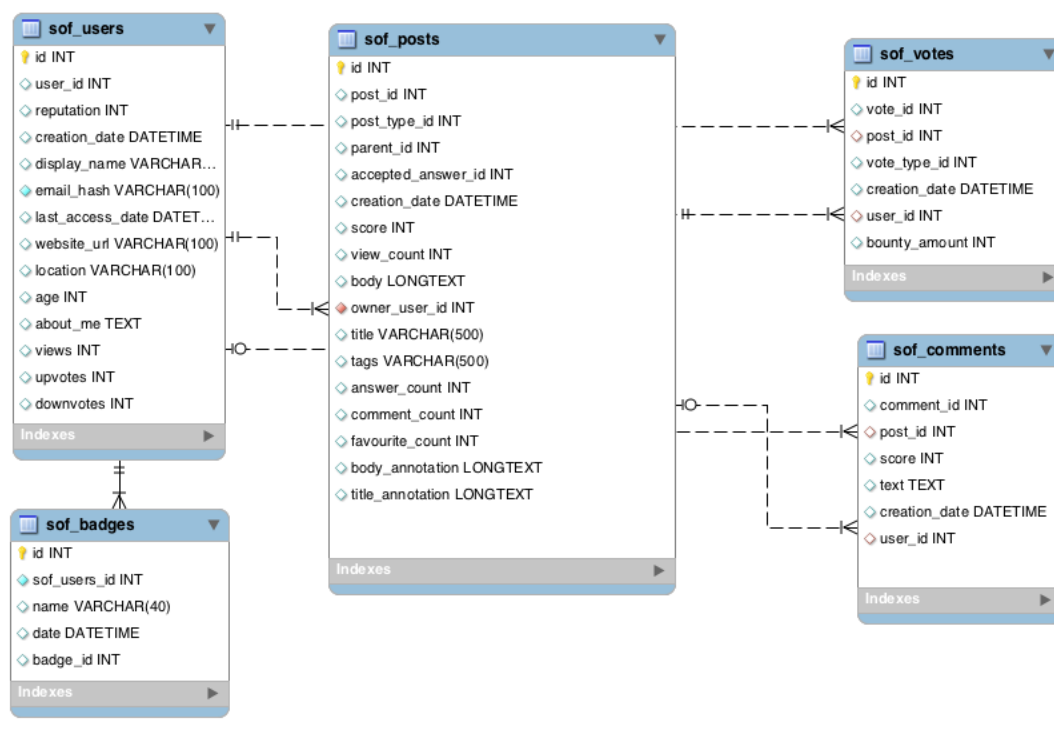


Figure 4.2: ER diagram of StackOverflow dataset

4.2.4.2 Reddit Data Mining

Reddit has an API that allows for the retrieval of posts from particular subreddits. For this experiment 15 programming related subreddits were chosen that corresponded to the top 10 tags of StackOverflow. These subreddits are – Java, PHP, Python, Javascript, Ruby, C++, C#, Perl, Programming, Learnprogramming and Webdev.

The PRAW⁴ (Python Reddit API Wrapper) library was used to get the posts, comments, votes, users, flairs, etc. information from every subreddit. The PRAW library dealt with

⁴<https://github.com/praw-dev/praw>

the API bandwidth and call limits. The JSON file was parsed by the library and stored in the database.

The main limitation of the Reddit API was that it did not give information after 200 pages and each page contained only 25 posts. So every subreddit only provided a limited number of posts. The Reddit API did not provide complete datasets in a particular subreddit like StackOverflow. Hence, the Reddit dataset and user profile was incomplete.

Figure 4.3 shows Reddit dataset schema. This was built to store the data temporarily in MySQL database.

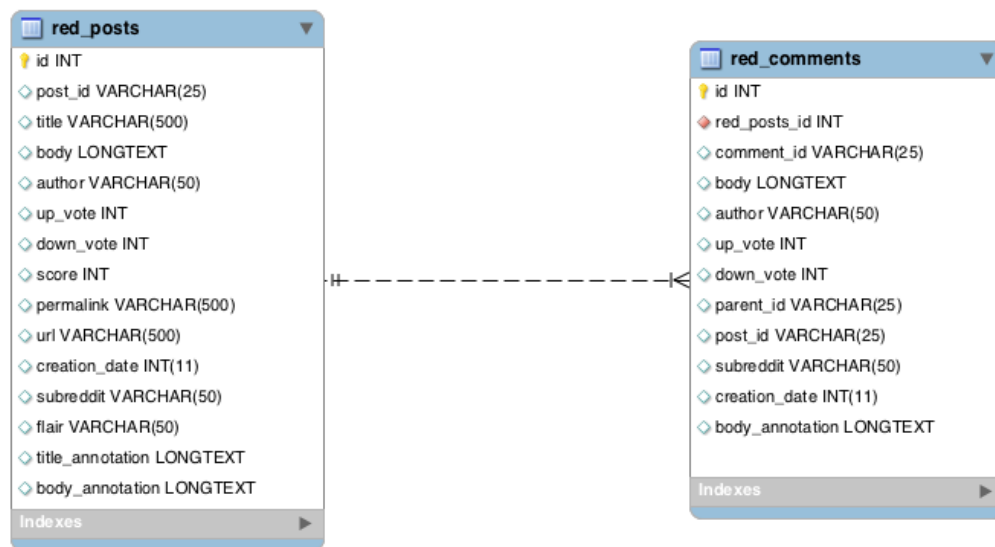


Figure 4.3: ER diagram of Reddit dataset

4.2.4.3 Data Structuring using RDF

StackOverflow and Reddit data were collected, encoded in Unicode format and stored in the MySQL database after mining. Some posts contained snippets of code and the code was stored as text in the database. The StackOverflow dataset consisted of 15 million questions, 28 million answers, 87 million votes and 2 million users. The whole dataset was too big for the work machine to cope with. To make the dataset manageable and still keeping it complete within the community, the top 10 tags were chosen. This trimmed down the dataset into 7.2 million questions, 12.9 million answers, 63 million votes and 2 million users. This didn't affect the community structure because the community data were complete and all the data for top 10 tags were complete. When choosing the data all the related tags and synonyms of the tags were also included to keep related communities together. The complete dataset of the tags was turned into RDF.

Reddit data consisted of 190 thousand posts, 0.39 million comments and 172 thousand users collected. In comparison to StackOverflow the Reddit dataset was smaller. This dataset was converted into RDF as well.

4.2.4.3.1 Generating RDF As mentioned in the earlier section, FOAF ontology was used to describe the users and their profile information. Both StackOverflow and Reddit only shows basic user profile information due to privacy reasons.

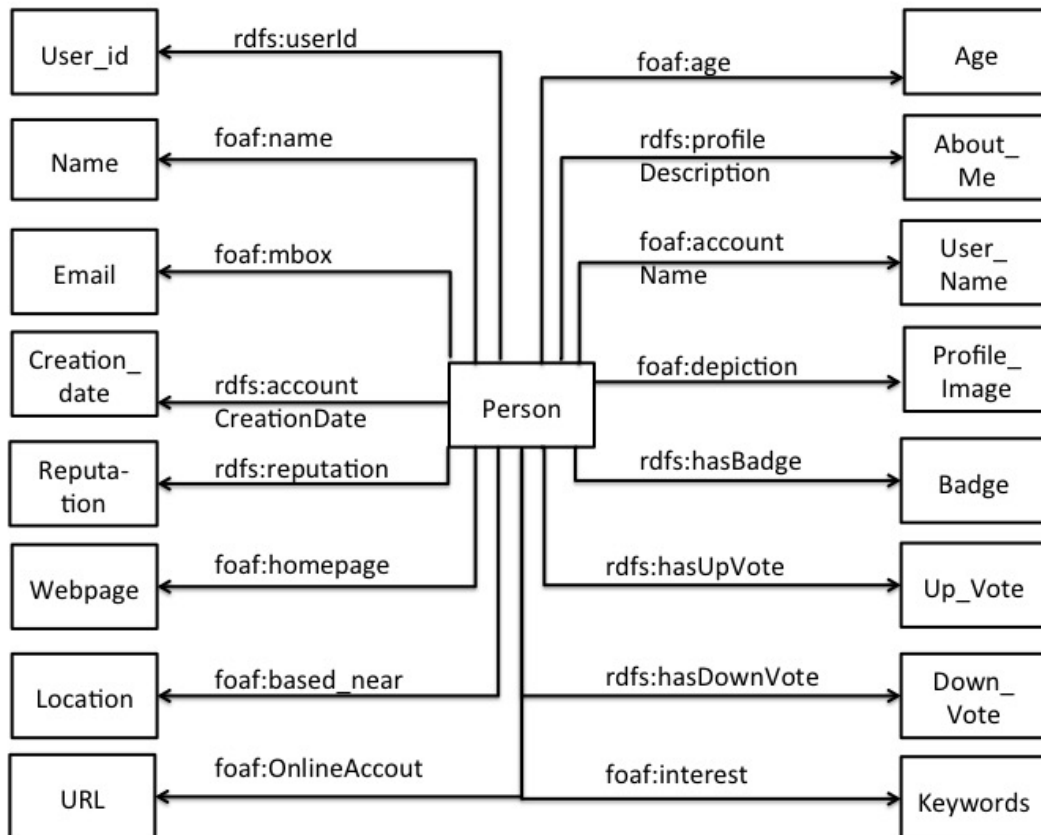


Figure 4.4: RDF schema of User's profile

An example of a simple user profile information is as follows:

```

<foaf:Person>
  <foaf:name> Geoff Dalgas </foaf:name>
  <foaf:mbox_sha1sum> b437f461b3fd27387c5d8ab47a293d35 </foaf:mbox_sha1sum>
  <foaf:based_near> Corvallis, OR </foaf:based_near>
  <foaf:age> 35 </foaf:age>
  <foaf:OnlineAccount> http://stackoverflow.com/users/2/geoff-dalgas
  </foaf:OnlineAccount>
</foaf:Person>
  
```

Similarly, the posts created by users, the questions and answers are described using SIOC and DC ontology. The post is linked with the users' profile data by using the same URIs for the user.

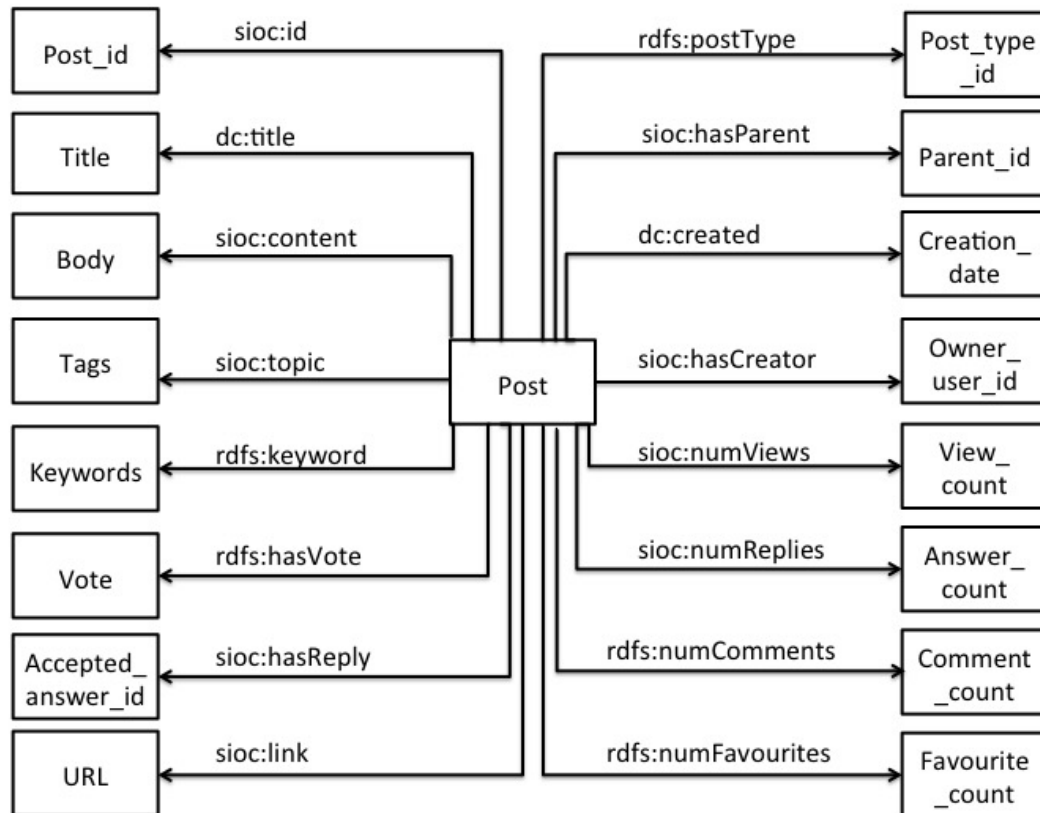


Figure 4.5: RDF schema of StackOverflow posts

An example of a simple posts information is as follows:

```

<sioc:Post rdf:about=" http://stackoverflow.com/questions/89228/calling-an-external
-command-in-python">
  <dcterms:title>Calling an external command in Python</dcterms:title>
  <dcterms:created> 2008-09-18T21:42:52.667 </dcterms:created>
  <sioc:has_container rdf:resource=" http://stackoverflow.com/questions/tagged
/python"/>
  <sioc:has_creator>
    <sioc:UserAccount rdf:about=" http://stackoverflow.com/users/170339/bludger "
      rdfs:label="bludger"> </sioc:UserAccount>
  </sioc:has_creator>
  <sioc:content>How can I call an external command in Python</sioc:content>
  <sioc:topic rdfs:label="python" rdf:resource=" http://stackoverflow.com
/questions/tagged/python"/>
  <sioc:topic rdfs:label="command" rdf:resource=" http://stackoverflow.com

```

```

/questions/tagged/command"/>
<sioc:has_reply>
  <sioc:Post rdf:about=" http://stackoverflow.com/a/89243/1313327">
    <sioc:content>Look at the subprocess module in the stdlib: from
    subprocess import call call(["ls", "-l"]) The advantage of subprocess
    vs system is that it is more flexible (you can get the stdout,
    stderr, the "real" status code, better error handling, etc...). I
    think os.system is deprecated, too, or will be: http://docs.python.org
    /library/subprocess.html#replacing-older-functions
    -with-the-subprocess-module For quick/dirty/one time scripts, os.system
    is enough, though.</sioc:content>
    <dcterms:created>2008-09-18T23:42:52.667</dcterms:created>
    <sioc:has_creator>
      <sioc:UserAccount rdf:about="http://stackoverflow.com/users/
      11465/david-cournapeau" rdfs:label=" david-cournapeau ">
        </sioc:UserAccount>
      </sioc:has_creator>
    </sioc:Post>
  </sioc:has_reply>
</sioc:Post>

```

As seen from the figures 4.4, 4.5 and 4.6 that are many fields in StackOverflow and Reddit data that cannot be defined by FOAF, SIOC or DC ontologies. For example, fields like the up votes and the down votes, they do not have equivalent properties in these ontologies. These are some limitations to the FOAF and SIOC ontology. In these special cases, RDF schema was defined to add more vocabulary to describe the upvotes and downvotes given to the questions, answers and comments. Three properties were described in the RDFS. They are - :hasVote, :hasUpVote, :hasDownVote

The properties are self-explanatory. :hasVote describes the total number of vote a post has. It can be a positive or negative integer. :hasUpVote shows the number of up votes a post has and similarly number of down votes was shown by :hasDownVotes property.

The other issue with generating RDF was to give both StackOverflow and Reddit data a common structure. The datasets mined from these websites were different. StackOverflow data consisted of questions and answers. Reddit data consisted of posts and comments. The comments also had children comments. The Reddit posts were of two types - text posts that were questions; and information and link posts that linked to external sources.

The issue was resolved by considering all questions, answers and comments as posts. These posts were of two types to differentiate their purpose.

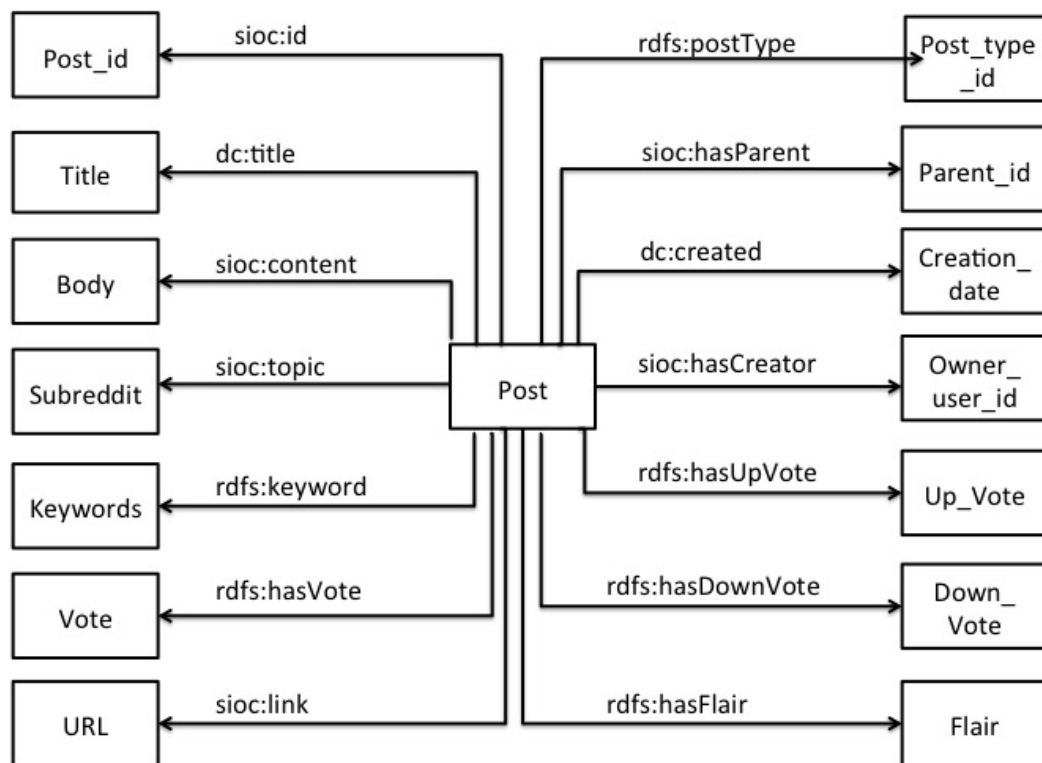


Figure 4.6: RDF schema of Reddit posts

Type 1 was for StackOverflow questions and Reddit main posts. Type 2 was for StackOverflow answers and Reddit comments (parent and children comments).

Each post had a parent post associated with it. The Type 1 post had no parent posts. Type 2 posts that consisted of answers and main comments had their corresponding parent posts as parents. The children comment posts in Reddit had the main comment post as their parent post to maintain the thread structure of the conversation. Figure 4.7 explains this relationship.

The other issues were that the mined data from both websites was not homogeneous (e.g. StackOverflow had badges and reputations whereas Reddit has karma and trophies). This issue was resolved by adding RDF properties for every different information. If the data was available, then it was added otherwise it was ignored. One important example for such issue is if StackOverflow question has ‘accepted answer’ property that contains the id of the answers that the user accepted to solve their problem. Reddit doesn’t contain such data, instead it has ‘flair’ data that give extra information about the post.

4.2.4.3.2 Database All the RDF created from StackOverflow and Reddit datasets was stored in the Stardog database. The RDF schema file containing all the newly created properties was also stored in the Stardog database. There were more than 45

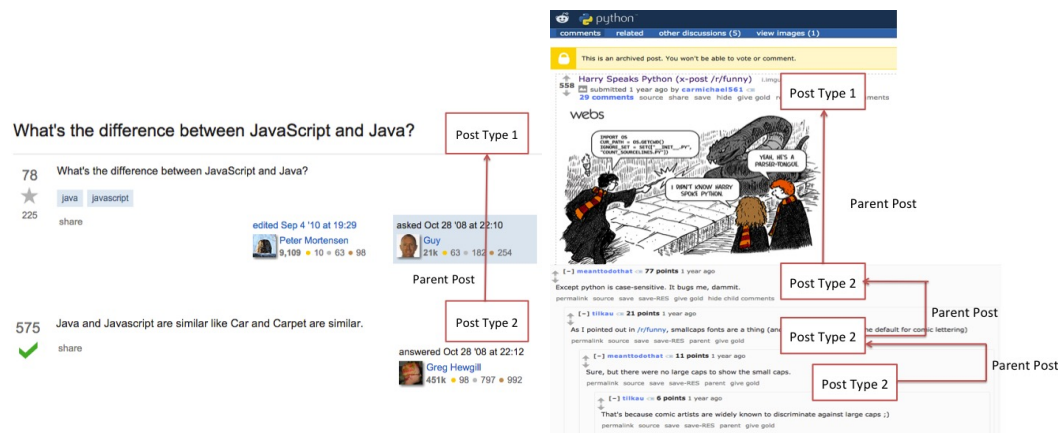


Figure 4.7: Showing parent-child relationships in 2 different types of posts in StackOverflow and Reddit.

million triples that contain all the information about the posts and more than 3.2 million FOAF profiles in the dataset. 15 thousand of them are the unique annotated keywords that are linked to the DBpedia dataset and 6 thousand are linked to Drupal dataset.

4.2.4.4 Keyword Annotation and Linking

The StackOverflow dataset is sparsely annotated by user-generated tags and it is not linked with any other datasets. When a user posts a question on the website, they add tags to it. The tags help to categorize it into different topics and show it on the different tags page and notify users subscribed to that tag. The answers on the other hand do not have any tags. They inherit the tags from the questions. During data collection, question tag was added to the answers by the data mining script.

The Reddit dataset has no tags associated with posts. During the data mining process the name of the subreddit was added as the tag to both posts and comments. This worked for language specific subreddits like Python and PHP, but did not work for general programming subreddit like 'learnprogramming'.

The absence of tags in Reddit posts caused problems in finding the topic and category of the questions, answers and posts. The main topics in the text of questions and answers were not clearly stated. The topics were ambiguous and not linked to any vocabulary or concepts. This was one of the problems resolved by the Suman system by doing keyword disambiguation using Wikipedia Miner and OpenCalais.

The questions, answers, comments and tags data were annotated with the links from Wikipedia datasets and OpenCalais datasets to resolve names and topics ambiguity. The returned keywords were further transformed into RDF. By using the links to the matched topics StackOverflow and Reddit data was linked to the Wikipedia and OpenCalais dataset.

4.2.4.4.1 Wikipedia Miner: Wikipedia Miner is used to annotate posts from top 10 tags of StackOverflow and Reddit. The service returns the text with keywords associated with it. The service accepts simple text or HTML and returns different file formats like XML, JSON, etc. The service is configurable, one can specify the density of links to be added and the level of accuracy required from the service. For this experiment the threshold score was 0.75. Any keyword match below this score was ignored to maintain the quality of the keywords. The Wikipedia categories are also extracted with the keywords.

Figure 4.8 is an example annotated text of a question and answer posted.



Figure 4.8: Wikipedia Miner annotation example

4.2.4.4.2 OpenCalais: OpenCalais is another web service used to annotate the StackOverflow and Reddit posts with the OpenCalais dataset. An example annotation of text using OpenCalais is below.

Consider this sample question: "Does Python have a ternary conditional operator? If not available, is it possible to simulate one concisely using other language constructs?"

This question is annotated by OpenCalais and following annotations are returned by this service:

```
<Programming count="1" relevance="0.649" normalized="Python (programming language)">Python</Programming>
```

```
<Programming count="1" relevance="0.868" normalized="Software Engineering">Software engineering </Programming>
```

```

<SocialTags>
  <SocialTag importance="2"> Conditional
<originalValue>Conditional (programming)</originalValue>
  </SocialTag>
  <SocialTag importance="2"> Python
  <originalValue>Python (programming language)</originalValue>
  </SocialTag>
  <SocialTag importance="2"> C
  <originalValue>C (programming language)</originalValue>
  </SocialTag>
  <SocialTag importance="2"> Ternary operation
  <originalValue>Ternary operation</originalValue>
  </SocialTag>
  <SocialTag importance="1"> Software engineering
  <originalValue>Software engineering</originalValue>
  </SocialTag>
  <SocialTag importance="1"> Computing
  <originalValue>Computing</originalValue>
  </SocialTag>
  <SocialTag importance="1"> Computer programming
  <originalValue>Computer programming</originalValue>
  </SocialTag>
</SocialTags>

```

As seen from the above snippet, the OpenCalais service finds the keywords and matches it with its own taxonomy or vocabulary. It assigns the importance to the disambiguated tag. It also adds an entity relevance score to each keyword. Similar to Wikipedia Miner, it is configurable. OpenCalais accepts a different format of input and provide different formats of outputs. It accepts threshold for entity relevance score and only provide matches above the threshold limit. A similar threshold limit of 0.75 was set for finding keywords. The entity relevance score is comparable across the input level so its relevancy is determined at collection level, not just at document level (Reuters, 2008).

OpenCalais adds SocialTag to extract predefined structured information. This is another way to extract the context of the content. A topic extracted by ‘Categorization’ with a score higher than 0.6 will also be extracted as a SocialTag. If its score is higher than 0.8, its importance as a SocialTag will be set to 1. If the score is between 0.6 and 0.8 its importance is set to 2. This also allows to add categories to the dataset. The keywords are linked to the OpenCalais dataset.

All the keywords annotated using Wikipedia Miner and OpenCalais were linked with DBpedia and OpenCalais dataset. There were more than 15 thousand unique annotated keywords that were linked to the DBpedia dataset and 6 thousand were linked to OpenCalais dataset.

4.2.4.5 Information Retrieval

The unanswered questions from StackOverflow and Reddit were used as a query to find similar questions with answers. This was done to test the hypothesis to see if the resulted answers could be used to answer the original unanswered questions. The Suman system also recommended experts who could answer the questions.

The Suman algorithm as discussed earlier was used with StackOverflow and Reddit dataset. Unanswered questions from StackOverflow and Reddit were used to test the system. The system searches for answers to an unanswered question. So a query consisted of an unanswered question and the keywords associated with it. The final result consisted of ranked list of similar questions with answers that could potentially answer the original unanswered query question. The Suman algorithm also recommended a list of experts that were best suited to answer the unanswered question.

4.2.5 Implementation Evaluation

The Suman system needs to be evaluated to see if it satisfies all the requirements mentioned in the problem investigation section. The system also needs to meet all the research challenges and it should be tested to see it does what it intends to do.

The Suman system is evaluated on these criteria:

1. Uses PSNs data. Collects it and harvests it.
2. Solves name entity disambiguation problem.
3. Uses Linked Data and Semantic Web technologies.
4. Uses semantic search to find answers to unanswered questions.
5. Searches for experts that could answer the unanswered questions.

The Suman system fulfills the first criteria because it uses StackOverflow and Reddit datasets as its knowledgebase. It uses Wikipedia Miner and OpenCalais to solve Name Entity Disambiguation, hence fulfilling the second criteria. All the questions and answers have semantics added to it adding extra meaning to the objects. These keywords are

linked and weighted. It also links the data to Wikipedia and Drupal datasets thus fulfilling the third criteria for linking the datasets.

The Suman system is tested using unanswered questions from StackOverflow and Reddit to fulfill and the last two criteria. The details of these tests are given below.

4.2.5.1 Keyword Annotation and Categories Analysis

This section gives an overview of the added annotation to the questions and answers by the Suman system. This is done by doing a statistical analysis of all the data in StackOverflow and Reddit that has been semantically enriched.

The keywords were added to questions and answers of the top 10 tags in StackOverflow datasets.

The Suman system annotated the question and answers from the DBpedia and OpenCalais websites. The annotation procedure adds keywords to the keywords table with the degree of confidence. The advantage of adding keywords is that in StackOverflow website users are restricted to the already existing list of keywords and then need certain reputation points to add a brand new keyword. No such restriction existed in the Suman system. Also, the system annotated all the occurrence of a keyword thus giving it a higher degree of confidence. If a keyword occurred multiple times in a question and answer, then it was one of the important keywords to describe the post.

The analysis of keyword annotation shows that the system added more than 20 million keywords to the posts and more than 15 thousand were unique keywords. These unique keywords were repeated multiple times within the question and answers. On average 4.87 keywords were added to each post.

To maintain the quality of the keywords, the degree of confidence was kept at 75%. Any keyword with a lower score was ignored by the system. This reduced the number of overall added keywords. In the beginning when the confidence score was kept at 50%, the analysis of 1000 posts showed that the average number of keywords was 9.21 keywords per post but it also annotated the not technical texts. Then the value of the degree of confidence was increased and the keyword category was limited to the technical terms.

The keywords were also categorized using the DBpedia categories. Every keyword had a list of categories it associated with and the list was separate from the keywords. The dataset had more than 600 categories in total and each keyword had 3.4 categories in average. The categories didn't have a degree of confidence level because the list was directly taken from the DBpedia website.

Similar to StackOverflow, Reddit posts and comments are annotated too. The system added 2.6 million keywords to the posts and comments. 9.3 thousand were unique

keywords, i.e. these unique keywords were repeated multiple times within the question and answers. On average 3.31 keywords were added to each post. Just like before the degree of confidence is kept greater than 50%.

The keywords were also categorized like StackOverflow data. The dataset had more than 500 categories in total and each keyword had 3.2 categories in average. The categories didn't have a degree of confidence level because the list was directly taken from the DBpedia website.

4.2.5.2 Answers and Experts Analysis

The Suman system searches for answers for the unanswered questions in the database and gives each answer a confidence score between 0 and 10. The system was tested using the unanswered questions from the month of July 2014. There were 20,326 unanswered questions in the top 10 tags and the system searched for relevant answers with confidence score more than 75% for 13,209 questions. 23.62% of unanswered questions had one or more answers with confidence score of 85% and 82.27% of unanswered question has a confidence score of more than 50%.

Answers' confidence score	Number of questions	Percentage of questions
>90%	2623	12.9%
81-90%	3104	15.27%
71-80%	5845	28.75%
61-70%	3804	18.71%
50-60%	1348	6.63%

Table 4.1: StackOverflow: Percentage of questions with answers with confidence score

In the search result any answer with confidence score less than 50% is rejected and only the top 3 answers are shown. The table 4.1 shows all the answers (without the limit of 3) available for each range of confidence score for the unanswered questions. For the evaluation of answers the algorithm was changed to show the lower rating answers and the number of questions with one or more answer changed to staggering 97.72%.

The search result also showed the list of experts recommended to answer each unanswered question. The expert list followed the same pattern as the answers. It only showed the top 5 experts with confidence score higher than 50%. If the question didn't have any answer with confidence score greater than 50% even then the system would show a list of recommended experts. This list was also restricted to the experts with confidence score of greater than 50%.

There were 20067 questions with at least one recommended expert with a confidence score greater than 50%. This equates to 98.73% of the unanswered questions on StackOverflow.

Experts with confidence score	Number of Experts (%)
>90%	14.74%
81-90%	16.31%
71-80%	27.66%
61-70%	22.87%
50-60%	13.92%

Table 4.2: StackOverflow: Percentage of experts with confidence score

The Suman system search algorithm was used to search for unanswered questions in Reddit for the month of June 2014. There were 1381 posts with no comments, these were considered as unanswered questions for analysis. The Suman system searched and found one or more answers for 84.5% posts with a degree of confidence more than 50% and 7.89% has answers with confidence score more than 85%.

A quick glance at the results showed that most of the answers 96.37% came from StackOverflow.

Answers' confidence score	Number of questions	Percentage of questions
>90%	48	3.47%
81-90%	81	5.86%
71-80%	425	30.77%
61-70%	342	24.76%
50-60%	271	19.62%

Table 4.3: Reddit: Percentage of posts with answers and their confidence score

The top 3 answers shown as the search result. When the limit of confidence score is removed, 99.79% of unanswered questions has one or more answers.

Again, similar to StackOverflow, top 5 experts are recommended to each question. 98.82% of questions has one or more recommend expert. The table 4.4 lists the percentage of questions and list of recommended experts.

Experts with confidence score	Number of Experts (%)
>90%	12.65%
81-90%	17.23%
71-80%	24.38%
61-70%	26.29%
50-60%	11.76%

Table 4.4: Reddit: Percentage of experts with confidence score

One thing was noticed in all the search results and expert recommendations of Reddit is that most of the answers and experts were recommended from StackOverflow. The reason behind this could be that the StackOverflow website data is huge compared to Reddit dataset (see tables 3.1 and 3.8). So the users in StackOverflow website have

more keywords and reputation point associated with them. Similarly, the answers are affected too.

Also, since the StackOverflow and Reddit databases were combined when it was converted to RDF. It wasn't possible to test the quality of Reddit answers and users' expertise by completely disabling the StackOverflow data and checking for quality of answers and experts.

4.2.5.3 Expert Recommendation Analysis

The Suman system used the reputations and karma points to search experts based on the keywords and recommend them as experts for the unanswered questions. The system only picks users that have answered any questions and have at least one positive vote for the answer. StackOverflow and Reddit users are merged together in the expert database so they can't be analyzed separately.

There are 3.2 million users in the database that can be considered experts and 20 thousand keywords. The experts are recommended based on individual keywords and also set of keywords taken together. The keywords also have categories, so the experts can also be recommended based on categories. The recommendation algorithm not only picks the top expert for each keyword, it also picks them based on a set of keywords and categories.

For example, according to the StackOverflow website the top user or an expert of C# is Jon Skeet with more than eighty thousand reputation points and Python is Alex Martelli with more than nineteen thousand reputation points. These users appear on the individual pages of the tags as the top users and without the tag disambiguation they only appear as an expert on a particulate tag, not the joint concept of the topic.

Tag	Top User with reputation point
C#	Jon Skeet (80.6k)
Java	Jon Skeet (39.7k)
Python	Alex Martelli (19.8k)
PHP	Pekka (9k)
Javascript	CMS (12.3k)

Table 4.5: Top users of top tags in StackOverflow

When the tags are disambiguated and the keywords are matched to the topics, both Java and C# is categorized at the Object Oriented programming language and here Jon Skeet is considered as an expert in the whole area with more than one hundred and twenty thousand reputation points. Similarly, when the programming languages are further categorized as server side script ion language with Python, PHP and Perl as main languages, Alex Martelli is considered as an expert and the user CMS is an expert

in the client side languages such as Java and AJAX with twelve thousand reputation points.

Disambiguated Keywords	Top Expert with reputation point
Object Oriented programming (C#, Java)	Jon Skeet (120.3k)
Programming language(C#, Java, Python)	Jon Skeet (120.7k)
Server side Scripting language (Python, PHP, Perl)	Alex Martelli (20.2k)
Clientside Scripting language (Javascript, AJAX)	CMS (12.3k)

Table 4.6: Top users of top disambiguated topics in Suman system.

This becomes more complicated when the set of tags are considered. For example, when Java and MySQL are used together then the top expert is JB Nizet with 9k reputation point. It is none of the original experts. Again, when another set if used Java, MySQL, Database and Lucene, then Mauricio Scheffer is considered the expert with 4.3k reputation points.

The Suman system provides more flexibility and diversity in generating the expert lists that are not originally present in the main websites. It also provides experts across the domain without any restriction of website and can help in better search and discovery of experts and information.

4.2.5.4 Linked Data Graph Analysis

The StackOverflow and Reddit data are annotated, converted into triples and linked with DBpedia and OpenCalais datasets. Wikipedia Miner and OpenCalais applications are used to annotate and link it to the Linked Data cloud.

There are more than 45 million triples that contain all the information about the posts and users. 15 thousand of them are the unique annotated keywords that are linked to the DBpedia dataset and 6 thousand are linked to OpenCalais dataset.

User data is converted using FOAF and there are more than 3.2 million FOAF profiles in the dataset.

<a href="http://spotlight.dbpedia.org/rest/document/?text=<p>Is there a standard way for a Web Server to determine what time zone offset a user is in? </p> <p>From a <code>HTTP</code> header, or part of the user-agent description perhaps?</p>#offset_33_43">http://spotlight.dbpedia.org/rest/document/?text=<p>Is there a standard way for a Web Server to determine what time zone offset a user is in? </p> <p>From a <code>HTTP</code> header, or part of the user-agent description perhaps?</p>#offset_33_43	http://www.w3.org/2005/11/its/rdf#disambigIdentRef	http://dbpedia.org/resource/Web_server
<a href="http://spotlight.dbpedia.org/rest/document/?text=<p>Is there a standard way for a Web Server to determine what time zone offset a user is in? </p> <p>From a <code>HTTP</code> header, or part of the user-agent description perhaps?</p>#offset_33_43">http://spotlight.dbpedia.org/rest/document/?text=<p>Is there a standard way for a Web Server to determine what time zone offset a user is in? </p> <p>From a <code>HTTP</code> header, or part of the user-agent description perhaps?</p>#offset_33_43	http://www.w3.org/1999/02/22-rdf-syntax-ns#type	http://nlp2rdf.lod2.eu/schema/string/OffsetBasedString
<a href="http://spotlight.dbpedia.org/rest/document/?text=<p>Is there a standard way for a Web Server to determine what time zone offset a user is in? </p> <p>From a <code>HTTP</code> header, or part of the user-agent description perhaps?</p>#offset_115_119">http://spotlight.dbpedia.org/rest/document/?text=<p>Is there a standard way for a Web Server to determine what time zone offset a user is in? </p> <p>From a <code>HTTP</code> header, or part of the user-agent description perhaps?</p>#offset_115_119	http://www.w3.org/2005/11/its/rdf#disambigIdentRef	http://dbpedia.org/resource/Hypertext_Transfer_Protocol

Figure 4.9: An example to keywords N-Triple linked to DBpedia

It is concluded by the all the analysis done that the Suman system design meets the requirement to solve DP1 and DP2. It fulfills all the design goals and meet the design requirements. The validity of the keywords and the answers are checked in the next chapter as knowledge questions and they discussed in details as KQ3, KQ4 and KQ5.

Chapter 5

The Suman System Evaluation

The Suman system has been designed to answer the design problems DP1 and DP2. It is a prototype system build to answer RQ1. To validate that the system works, it is required to evaluate the system. This raises other knowledge questions KQ3, KQ4 and KQ5. Knowledge questions are the descriptive questions that can used to answer descriptive questions about the artifacts, in this case the Suman system. They are mentioned in Chapter 1.2.

These knowledge questions use Design Science Empirical Cycle for problem solving. This chapter uses the design methodology discussed in (Wieringa, 2014).

5.1 Knowledge Problem Investigation

The main knowledge questioned discussed in this chapter are as follows:

KQ3 is to investigate if Linked Data technologies is useful in search and discovery of information in PSNs. This is tested by measuring if Linked Data technologies improves understanding the underlying meaning and concepts of questions and answers in PSNs.

KQ4 is to investigate how well the Suman system works and answers the unanswered questions in PSNs. The Suman system needs to be tested to make sure it did not have any bugs and it provided the correct output.

KQ5 is to investigate how good are the recommended experts and if they could answer the unanswered questions.

The Suman system algorithm generates the following information from the StackOverflow and Reddit dataset.

1. Keywords with a degree of confidence

2. Answers with rank and degree of confidence
3. Experts name with rank and degree of confidence

All three output is investigated by the three knowledge questions stated above.

The main goal of these research questions are to investigate how well the Suman system answers the unanswered question and answer the main research question (RQ1).

5.2 Research Design

The main object of study for knowledge questions is to investigate if Linked Data and Semantic Web technologies provide useful means to search for answers in PSNs. The other object of study is the search results generated by the Suman system. The primary goal is to study how well the answers provided by the Suman system can answer the unanswered questions. Similarly, the usefulness of the recommended experts needs to be measured too.

Research design for each knowledge question is discussed below.

5.2.1 Knowledge Question 3 (KQ3)

To answer the KQ3, usefulness of semantics added to the questions and answers needs to be measured in the context of PSNs. This can be done by designing a single case mechanism experiment. In this experiment, a single instance of the Suman system could be tested to measure the usefulness and accuracy of the keywords that describe the questions and answers.

The Suman system uses Semantic Web and Linked Data tools to generate all the keywords and based on the keyword graph finds the answers and the experts. So, testing the usefulness of the keywords could prove the usefulness of the tools.

The first output of the Suman system is the semantically enriched data related to the StackOverflow and Reddit datasets. These semantically enriched keywords could be tested to see how relevant they are and if they add value to the data. This could be tested by during a user experiment to rate how well the semantically enriched keywords describe the data and add value to the datasets.

The main component of this research design can be broken down accordingly:

- Object of Study: The Suman system generated keywords that are semantically enriched.

- Treatment specification: There are system generated keywords and user generated keywords. A user experiment could be conducted where participants are asked to rate these keywords.
- Measurement specification: The keywords are rated on the standard scale and it could be compared. A T-Test would be a good statistical test to compare the mean rating of the paired dependent variables.
- Inference: The higher rated keywords are considered good. If the system generated keywords are better that means the semantically enriching data adds value to the datasets.

For the convenience, this test will be called Keyword T-Test in the rest of the chapter. If the result of this test is moderately strong it could be inferred that the Linked Data and Semantic Web technologies add value to PSN data. And it could be concluded that it could lead to improving in search and discovery of answers to the unanswered questions.

5.2.2 Knowledge Question 4 (KQ4)

The second output of the Suman system are the search results, specifically answers with the rating. These results need to be evaluated to find its usefulness and if it's the right answer for the question.

Another user experiment could be conducted that would ask users to rate how well the answers provide a solution to the unanswered question. Experts in the programming language would be able to tell if the answer provided for the question can apply to the problem and if the users will be benefited or not with the answers. These the algorithm ratings could be compared to the user ratings to see if there is any correlation. If the correlation is strong, it could answer KQ4.

The main component of this research design can be broken down accordingly:

- Object of Study: The answers produced by the Suman system after conducting a search for unanswered questions
- Treatment specification: A user experiment could be conducted where users are asked to rate how well the answers searched by the Suman system can answer the unanswered question.
- Measurement specification: The answers could be rated on a standard scale. The answers also have a rating from the algorithm. A correlation test would be good to compare user rating and algorithm ratings.

- Inference: If there is a correlation between the user rating and algorithm rating, then it could be inferred that the Suman system works and it can search for answers for unanswered questions.

For the convenience, this test will be called Q&A correlation test in the rest of the chapter. If there is a strong positive correlation between the user ratings and the algorithm rating it could be concluded that the Suman system does perform well and it successfully finds answers for the unanswered questions in PSNs.

5.2.3 Knowledge Question 5 (KQ5)

The third output of the Suman system is the experts with ratings that could potentially answer the unanswered questions. An experiment needs to be conducted to measure the usefulness of the list of experts.

The main component of this research design can be broken down accordingly:

- Object of Study: List of recommended experts generated by the Suman system that could potentially answer the unanswered questions.
- Treatment specification: A user experiment could be done to show the user profile and expertise to participants and ask them to rate how good the user could answer the unanswered questions.
- Measurement specification: Users could be asked to rate the user's expertise. But this is not a good measurement because you cannot give full information about a user's expertise. A correlation test would be good to compare user rating and algorithm ratings.
- Inference: This experiment is hard to conduct and the expert ratings could not be fully inferred by the user ratings.

This test is hard to test in a user experiment setting. Similar user experiment as above is not appropriate to test the expert recommendation system because of the following reasons.

- The Suman system returns the list of recommended experts and the algorithm score. It doesn't provide with any other information.
- To test the recommended expert list, participants need to know the complete user profile of the experts on the list.

- We can't provide a complete profile of the experts to the participants (due to size and time constraints)
- Participants cannot accurately judge the expertise level of an expert by seeing their name and algorithm score.

Therefore the two user experiments are not appropriate for the expert list. So the evaluation of the expert generator part of the Suman system has been dropped. Only the keywords and answers are evaluated.

5.3 Research Design Validation

The results of empirical cycle are fallible. The designed system might not fully meet the goals of the stakeholders and answers to the knowledge questions might have limited validity. So, the designed system and answers produced must be justified. This is done by validation of the designed system in terms of stakeholders' goals and requirements and the validation of inferences in the empirical cycle.

The different experiments and analysis mentioned earlier should be justified to show they are the right approach and research design to answer the knowledge question.

5.3.1 Knowledge Question 3 (KQ3)

The keywords T-Test experiment proposed is the user experiment to test the usefulness of semantically enriched keywords. There are preexisting user generated tags and keywords in StackOverflow and Reddit data. The Suman system performs Name Entity Disambiguation and add more semantically enriched tags. These two tags could be compared to see which set adds meaning to the questions and answer and describe them accurately. A statistical T-Test could be done to test them. This is a valid statistical test that compares the means of two sets of dependant variables.

- Object of Study justification: Semantically generated keywords are the basic aspect of the Suman system. These are the first output of the system and the search algorithms are based on this. Testing the keywords are justified to test if the Suman system is built on a good base.
- Treatment specification justification: Participants who have experience in the field can be deemed as good judge of the usefulness of the keywords. A user experiment to test the usefulness and validity of the keywords are justified.
- Measurement specification justification: A T-Test is a standard statistical test to compare means of the ratings for the two pairs of dependent variables.

- Inference justification: If the effect size r is significant and the $p < 0.05$ then the inference made is justified.

The Keyword T-Test is the valid test to answer KQ3. The inference of the results can be justified to answer and make conclusion for the Suman system. This experiment fully answers KQ3.

5.3.2 Knowledge Question 4 (KQ4)

The Q&A Correlation test is to evaluate how well the answers found by the Suman system search algorithm can provide a solution to the unanswered questions. The correlation between the algorithm ratings and the participants' ratings could provide useful inference.

- Object of Study justification: Search result answers are being tested here. This is the main research question in this thesis. It is justified to test the quality of the answers searched by the Suman system.
- Treatment specification justification: Participants who have experience in the field can be deemed as good judge of the usefulness of the answers. A user experiment to test the usefulness and validity of the answers are justified.
- Measurement specification justification: A correlation test gives a good indication of how many users agree with algorithm ratings. The users rate the answers on the standard scale of 1-10. The algorithm rates the answers on the same scale. This measurement can give clear indication of the algorithm is working correctly or not.
- Inference justification: The stronger the correlation between the user ratings and the algorithm ratings, it could be inferred that algorithm is finding the right answers to the unanswered questions and doing what it intends to do. SO the inference made in this experiment is justified.

The Q&A Correlation test is the valid test to answer KQ4. The inference of the results can be justified to answer and make conclusion for the Suman system. This experiment fully answers KQ4.

5.3.3 Knowledge Question 5 (KQ5)

The third experiment is to test how good is the recommended experts list. These experts could potentially answer the unanswered question. They are also generated by the Suman system so it needs to be tested.

- Object of Study justification: The recommended lists of experts generated by the Suman system who could potentially answer the unanswered questions.
- Treatment specification justification: Participants who have experience in the field can be deemed as good judge of the usefulness of the experts once they see their past answers. But the problem arises because you can't show experts complete history to the participants. This experiment will be difficult to conduct.
- Measurement specification justification: Hard to measure the usefulness of expert with their partial profile.
- Inference justification: This experiment's result can only be partially met. One cannot make a full claim that the list of experts are suitable by seeing their partial profile and answer submission history.

The expert recommender test is not a valid test to answer KQ4. The inference of the results cannot be justified to answer and make conclusion for the Suman system. because it will not fully answer KQ5.

5.4 Research Execution

The Suman system has been evaluated to make sure the added keywords are meaningful and the answers provide a solution to the questions by doing two experiments.

The first experiment is the keywords T-test. In the keyword experiment, participants are given a questionnaire that has unanswered questions with two sets of keywords. One set is the already existing keywords from StackOverflow and Reddit and the other set is the system generated keywords. Participants are then asked to rate the quality of the keywords based on the question.

The Q&A experiment evaluates the answers. In this experiment, participants are shown unanswered question and a similar question and its answer that provides solutions to the unanswered question. The participants are asked to rate the quality of the answer and how effectively it answers and unanswered question.

The experts generated by the system were not tested. Participants in the evaluation experiment did not evaluate the Expert Finder application. There was no easy and simple way to provide a complete user profile of every user to the participants in the experiment. The user profile would consist of user's posts, badges, votes and reputation points. This is a huge data and there was no easy way to show it to the participants and then ask them to rate the expert recommender. So, this output was not tested and KQ5 is not answered.

The detail of all the steps taken to conduct the two research experiments are discussed in details below.

5.4.1 Calculating Sample Size

In both experiments, to build a proper statistical model to test the Suman system, it is vital to get the range of questions, answers, keywords and user response. It is quite important to get the right sample size of data. It is essential to test the range of quality of answers that can be applied to the entire population of data. A right subset of these data can be used as a sample and a correct statistical analysis can be done to calculate the effectiveness of the Suman system.

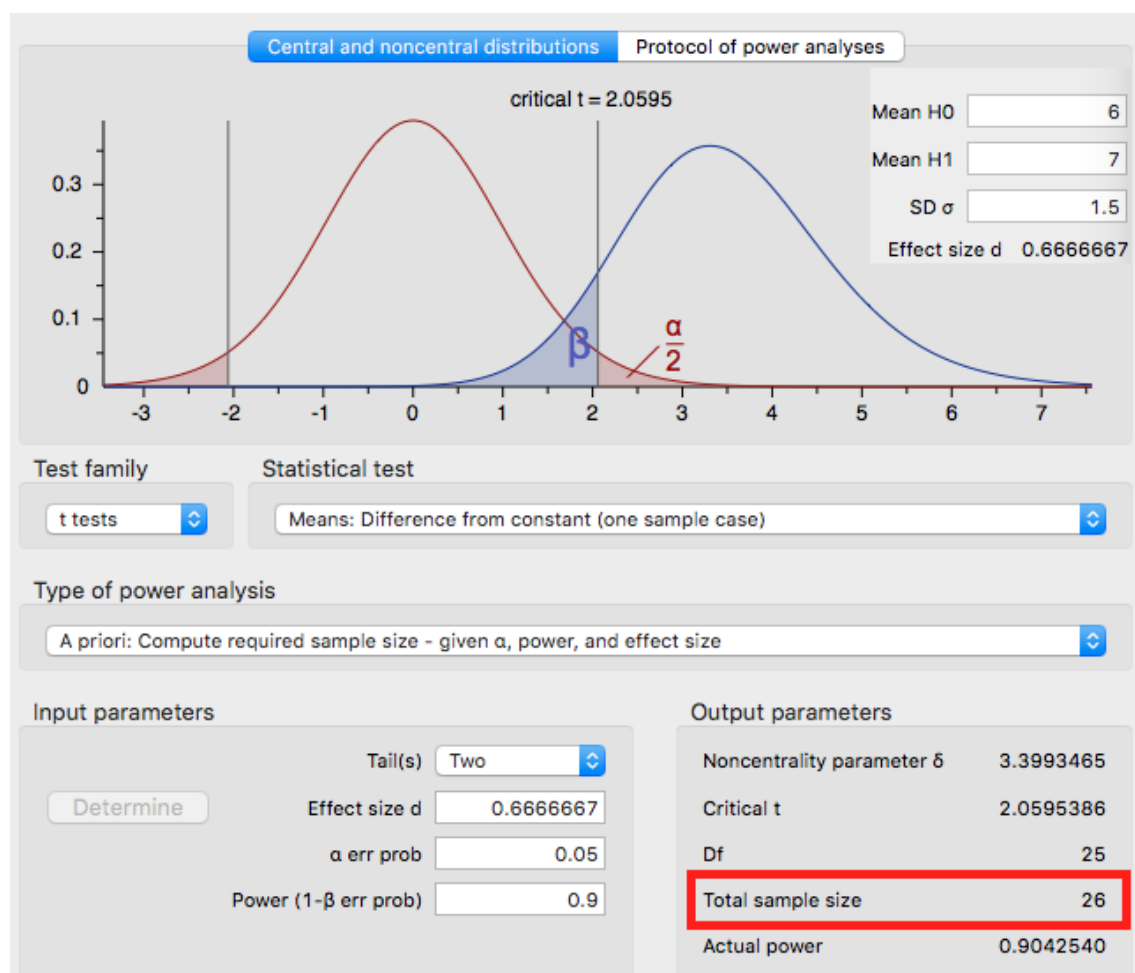


Figure 5.1: Calculating keyword experiment sample size using GPower

The sample size for both experiments (keywords and Q&A) are calculated using the G-Power software. All the values of alpha and beta errors, standard deviation, power, etc., are the standard value used in the research community.

- 1. **Keyword experiment:** Calculating the sample size using GPower gives the total sample size of 26. All the values used to calculate the sample size are shown in the figure 5.1
- 26 questions are required to get the proper sample size to represent the whole population of keywords in the Suman system.
- 2. **Q&A experiment:** Calculating the sample size using GPower gives the total sample size of 46. All the values used to calculate the sample size are shown in the figure 5.2

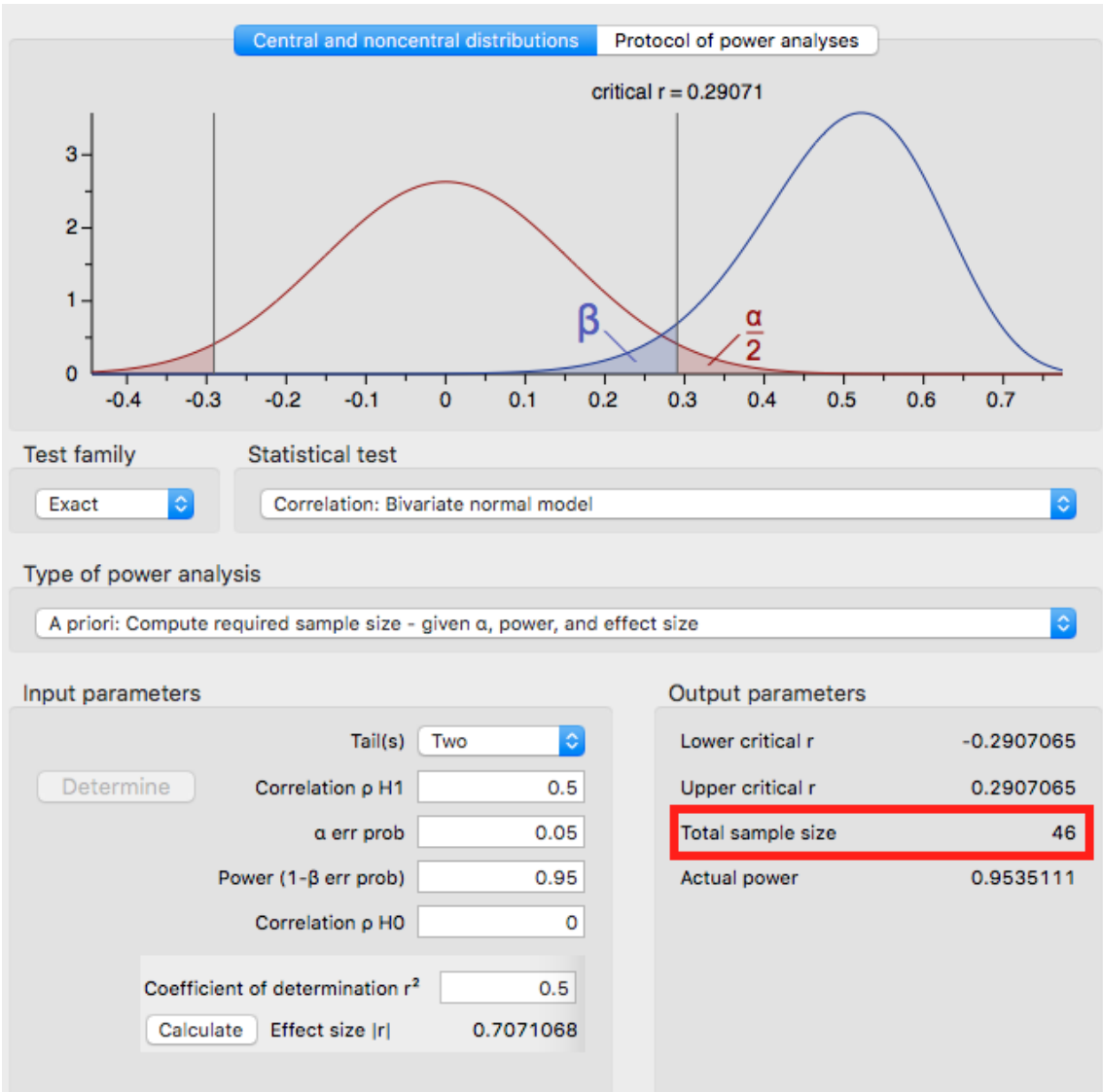


Figure 5.2: Calculating Q&A experiment sample size using GPower

46 answers are required to get the proper sample size to represent the whole population of answers in the Suman system.

3. **Participant size:** Calculating the sample size using GPower gives the total sample size of 20. All the values used to calculate the sample size are shown in the figure 5.3

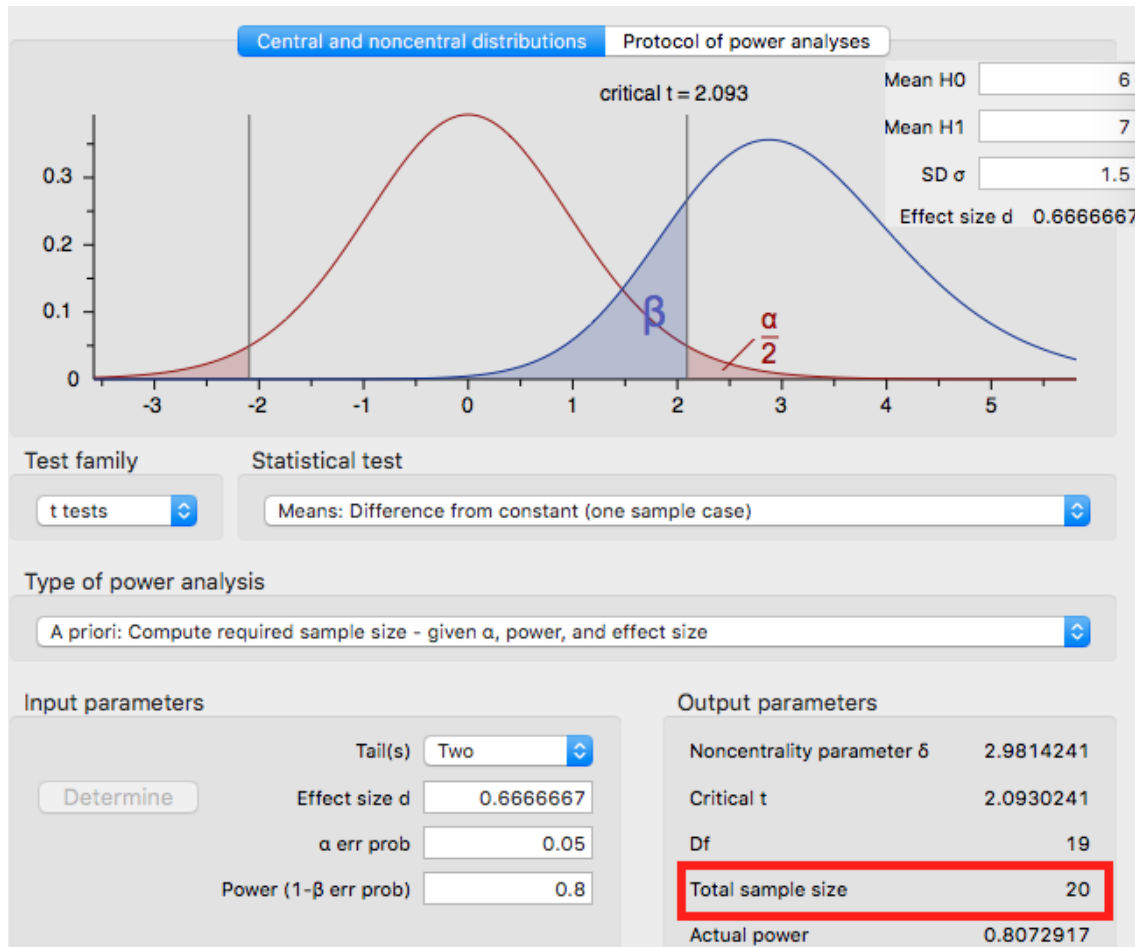


Figure 5.3: Calculating participants sample size using GPower

20 participants are required to get the proper sample size to properly evaluate the Suman system.

5.4.2 Selecting Questions

The next part of the experiment is selecting the right questions to ask the participants. A good range of sample questions must be chosen that represents all the different aspects of the dataset.

As calculated earlier, there needs to be 26 questions for the keywords experiment and 46 questions for the answers experiment. Each question requires 20 answers to get good and valid results.

There were certain factors that needed to be considered for the questions selections procedure.

1. First, the University's Ethics Committee says that an experiment cannot be longer than 1 hour.
2. The questions have to be short and easy to understand so the participants can finish the experiment in 1 hour.
3. The questions must not have any answers and must have keywords.

Due to time constraints the keyword experiment was split into two groups. Each group would consist of 20 people and answer 15 questions. And the Q&A experiment was split into three groups where two groups would answer 15 questions each and the third group would answer 16 questions.

Also, since the participants needed to answer programming related questions, they needed to be competent in programming and have a moderate expertise on the topic. This was determined by selecting the participants who had two or more years of experience in programming with the given programming language.

Java and Python programming language were chosen as these were the popular programming languages among the participants and also they are popular languages in both StackOverflow and Reddit. These were good language to get a wide variety of questions that met the criteria.

1. Topic - Questions were selected from different topics to give a wide range and variety to the subject. These were from web development, networking, mobile development, etc. Both Java and Python provide multiple and different areas where it could be used to program. This is relevant for both the keyword and Q&A experiments.
2. Difficulty - The selected questions should be of various levels of difficulties. The results could potentially provide some indication as to whether the Suman system can accommodate different levels of difficulty.

There were three levels of difficulty- easy, medium and hard. The difficulty of the question was determined by going through the content of books about learning programming languages. The beginning chapters taught easier topics, middle chapters were harder but still doable topics and later chapters are of more difficult topics. Based on this logic the unanswered questions were selected for both Java and Python. This will later be used to do a more complicated statistical analysis of the results.

3. **Quality** - The final criteria was about the quality of answers given to the unanswered questions. The algorithm that searches answers for questions gives them a rank and rate of confidence to show how good or bad the quality of the answers are. The experiment wants to see the correlation between the algorithms ratings and participant ratings. So, a wide range of answers were selected of different quality and they were- good, medium and bad.

The algorithm is designed to give each answer a score from 0 to 10. The answers that have score of higher than 7 is considered good, a score between 4 and 7 is considered medium and the score less than 4 is considered bad.

Based on all of the above criteria 15 questions from Python and 31 questions from Java were selected for participants. Since, same questions to test for answers could be used for keywords, the same 15 questions from both Python and Java were used for keywords experiments too.

There was no conflict to choose the same question because the keyword test was done with only questions and two sets of keywords. One set of keywords were already collected from the website and the other set were generated by the Suman system. Only the top 10 keywords were selected based on the highest rate of confidence. The Q&A experiment was done by using the same unanswered questions and a similar question with its answer.

5.4.3 Questionnaire Design

The next step of the keyword and Q&A experiments was to design the questionnaire to get user feedback to evaluate the Suman system. The questionnaire was created online for easy access for participants and also data collection was easy and University's survey portal known as iSurvey¹ was used to create it. The portal made it easier to add questions and make public pages for the questionnaire, get consent and collect data into the database.

For this experiment and questionnaire a rating scale was used to measure the response of the participants. The ratings are to measure the effectiveness of the output of the Suman system. This helps to understand and measure the usefulness of the whole application. Keyword experiment used the scale of 1 to 5 and the answer experiment used the scale of 1 to 10. This is discussed later in much more detail.

Three questionnaires were created, two for Java and one for Python. The first Java and Python questionnaire had 15 questions to rate the quality of answers and 15 questions with two sets of keywords asking participants to rate the quality of the keywords. The last Java questionnaire only evaluated answers, not the keywords.

¹<https://www.isurvey.soton.ac.uk/>

At first a pilot experiment was done with three participants to get feedback about the questionnaire and if it could be improved. The participants provided many helpful feedback. After getting their recommendation, the language of the questionnaire was made much more simpler and easier to understand. The questions were made bold so it was clearer and some grammatical mistakes were fixed. The participant instruction was added at the top of every question instead at the beginning of the questionnaire. Also, the question occurrence of keywords was randomized.

After making all the changes with the new and improved questionnaire, random participants were contacted to do the experiment. The only criteria for the participant selection where they had the knowledge of either Python or Java for more than two years. All the participants' information was anonymized. The participants were asked to not search for answers, but to use their own expertise to judge the quality of keywords and answers. In case they weren't sure about the answers then they were asked to use their best guess or select the middle value of the scale.

The details of each experiment are given below.

5.4.3.1 Keyword Evaluation

The participants were shown a question with the original keywords used in the website and asked to rate how well the keywords describe the question. Then the participants were shown the same question with the original plus the added keywords, but they were the top 10 keywords based on the rate of confidence of each keyword. Again, the participants were asked to rate how well the new sets of keywords describe the question. All the ratings were stored in the database.

Figure 5.4 is an example question with the instructions.

The keyword experiment used the rating from 1 to 5 where 1 is for very bad and 5 is for very good. A scale of 1 to 5 was chosen because the keywords are the basis of application, the Suman system uses the original keyword used in StackOverflow website and Reddit's subreddit name and other keywords are added to it. The difference in keywords is visible easily. Also, only the top 10 keywords are shown to the participants. This didn't make too many alterations to the list of keywords. Participants could easily see that in the use and quality of keywords. Also, the value given to each set of keywords was going to be compared to others in the T-Test, for this a small scale of 1-5 could give the results with a less error probability. So it was decided to use the scale from 1 to 5.

Instruction for question:

[Question ID : 696]

1. Read the question (section A)
2. Rate how well the keywords describe the question in your opinion (section B)

A. Java, console.readPassword adds extra line. How to delete it?

When i use `console.readPassword()` to read user passwords through console, there is always one line added to the console.

How to disable this behavior or **how to delete that extra line** (and move the cursor after the last character in the line before)? What escape character to use?

Thanks

B. Keywords: java, console, password

Q. Please rate how well the keywords describe the question in your opinion (section B)

Very bad	1	2	3	4	5	Very good
	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	

Figure 5.4: Sample keyword experiment question

In this survey the occurrence of question with original keywords and the added keywords was randomized so the participants could not guess the patterns at which these questions would appear. This step was taken to reduce participants' bias and was recommended by participants during the pilot program as the original questionnaire followed the same pattern of first asking the question with original keywords and second with the generated keywords.

In total 30 questions were asked to evaluate keywords and each question receive 20 responses. The statistical analysis is done in the next section.

5.4.3.2 Answer Evaluation

The Q&A experiment to evaluate the quality of answers consisted of the questionnaire where the participants were shown an unanswered question and then a similar question with an answer. This answer was expected to provide a solution for the first unanswered question. The participants were asked to rate the quality of the answer based on how well the answer could give a solution to the question. All the ratings were stored in the database.

Figure 5.5 provides an example of the question and participant instruction.

For the answers' evaluation the scale of 1 to 10 was used where 1 was very bad and 10 was very good. The scale from 1 to 10 was chosen because the algorithm that searches for the answers gives a rank and confidence rating to the answers that lies between 0

Similar Question:

B. Python: default/common way to read png images

I haven't found a standard way in Python to read images. Is there really none (because there are so many functions for so many custom stuff that I really wonder that there are no functions to read images)? Or what is it? (It should be available in the MacOSX standard installation and in most recent versions on Linux distributions.)

If there is none, what is the most common lib?

Many search results hint me to Python Imaging Library. If this is some well known Python-lib for reading images, why isn't it included in Python?

Answer:

C. No, there are no modules in the standard library for reading/writing/processing images directly. But the most common library might be [PIL \(Python Imaging Library\)](#). Many projects are not included in the standard library because they are 1) totally optional and 2) cannot be maintained by the few Python core developers.

Q. Please rate how well the similar question and answer (section B and C) provide a solution to the first question (section A).

Very bad	1	2	3	4	5	6	7	8	9	10	Very good
	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	

Instruction for question:

1. Read the question in both section A and B
2. Read section C which is the answer to B
3. Rate how well the answer in section C also answers the question in section A in your opinion
4. You are allowed to click on [any links](#) if you want.
5. Please don't Google the question as it might show you StackOverflow results and its rating might influence your rating.

A. Read text from PNG with standard lib

Is there a way to read text from a PNG-File in Python by using only the standard libraries Python provides?

Figure 5.5: Sample Q&A experiment question

and 10. This rank is equivalent to the participants' ratings and it could be used to do a Spearman correlation test to see if the participants agree with the ranking of the Suman system.

The questions that were selected had a range of difficulty from easy, medium and hard. The answers also had a range of quality from good, medium and bad. The participant's response would give a good estimation if they agree with the Suman system's rating or not. These different types of questions were randomized so the participants can't find the pattern in the difficulty and quality of question and answers.

In total there were 46 questions and answers tested. There were two questionnaires, one in Java and other in Python with 15 questions. The third questionnaire was in the Java with 16 questions. All the questions had 20 responses from the participants. The details of the response and statistical test are discussed in the next section.

5.5 Result Evaluation

After conducting the experiments with the participants their data was collected and stored in the University's iSurvey website. The website allows the data to be downloaded in an Excel sheet. The data were cleaned and structured, it was used to create a frequency distribution chart. The mean value was also calculated to do the statistical analysis.

These values are added to the SPSS software to do the statistical analysis. The details of all the analysis are in the following sections.

5.5.1 Dataset Frequency Distribution

A frequency distribution gives the frequency or count of all the data points for the sample. Here a general overview of the collected data is provided.

The diagram below describes the frequency distribution of all the ratings for the keyword experiment. It is a normal distribution.

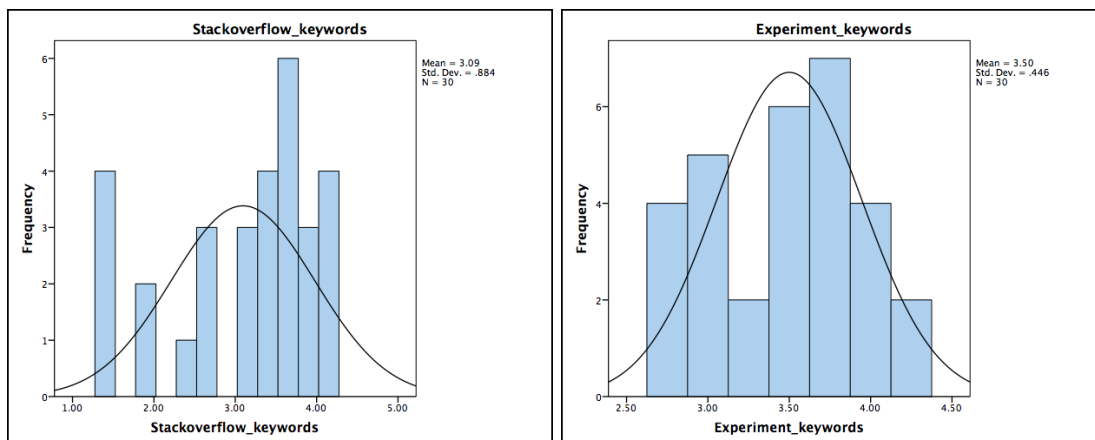


Figure 5.6: Keywords test frequency distribution diagram

The diagram is divided into two groups. The first diagram shows the frequency distribution of the ratings of the original keywords. The second diagram shows the frequency distribution of the keywords generated by the Suman system.

It is evident from the analysis that the system-generated keywords had higher ratings than the original keywords. Further statistical analysis of keywords will be done in the next section.

Similar analysis is done with the data collected from the Q&A experiment. The diagram below shows the frequency distribution of all the responses received from the 20 responses from the 46 questions.

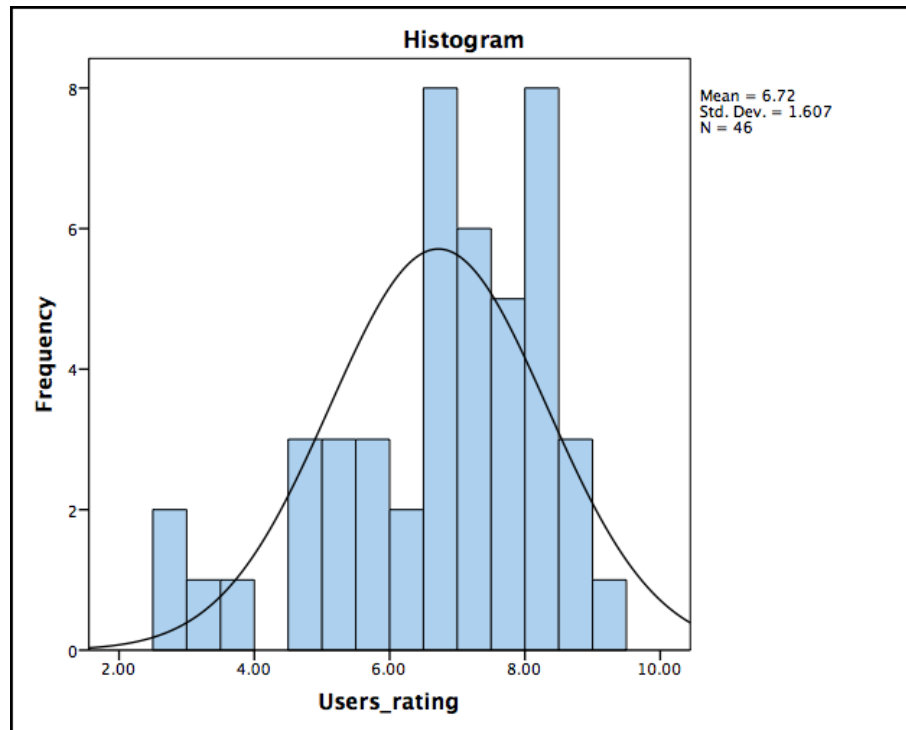


Figure 5.7: Q&A test frequency distribution diagram

The participants' response for the Q&A experiment is normally distributed too. More detailed statistical analysis of the answers is done in the following sections.

5.5.2 Keywords T-Test

When the data were collected, it had a set of keywords from StackOverflow and Reddit. The Suman system analyzed the text of the question and answers and added more keywords. It also added additional keywords to add broader and narrower categories to each data.

So, there was an original set of keywords and the added set of keywords. To evaluate the quality and usefulness of the added set of keywords the keyword experiment was designed. Participants were shown questions and asked to rate the quality of both sets of keywords based on how well they described the questions. So, there are ratings for how well the original sets of keywords describe a question and ratings for modified sets of keywords and how well they describe the same questions. Both the ratings are for the same question with little modification. They are a pair of the dependent variable.

Comparing the means of the ratings for the two pairs of dependent variables could provide the usefulness of a certain set of keywords. This could be calculated using a dependent T-Test. So, this particular statistical test was done to figure out if the generated sets of keywords were adding more value to the questions than the original

keywords. SPSS was used on the data collected during the experiment and Paired-Sampled T test was performed.

The analysis showed that on average the keywords generated by the Suman system were useful and described the question better (*Mean = 3.5, Standard Deviation = 0.44, Standard Error = 0.81*) than the original keywords (*Mean = 3.09, Standard Deviation = 0.88, Standard Error = 0.16*). There was a significant difference in the usefulness of generated keywords than the original keywords ($T(28) = -2.254, p = 0.032, r = 0.38$)

$$\text{Effect size } (r) = \sqrt{(t^2 / (t^2 + df))}$$

$$= \sqrt{((-2.254^2) / ((-2.254^2) + 29))}$$

$$= 0.38$$

T-Test

Paired Samples Statistics					
		Mean	N	Std. Deviation	Std. Error Mean
Pair 1	Stackoverflow_keywords	3.0950	30	.88409	.16141
	Experiment_keywords	3.5000	30	.44586	.08140

Figure 5.8: Keywords T-Test

Paired Samples Test									
		Paired Differences					t	df	Sig. (2-tailed)
		Mean	Std. Deviation	Std. Error Mean	95% Confidence Interval of the Difference				
					Lower	Upper			
Pair 1	Stackoverflow_keywords – Experiment_keywords	-.40500	.98422	.17969	-.77251	-.03749	-2.254	29	.032

Figure 5.9: Keyword T-Test showing confidence interval

Here $p < 0.5$ and the effect size $r = 0.38$ which is medium. The generated keywords add some benefit to the questions and answers. They improve in categorizing the topics of the questions and answers. The added keywords and categories could potentially improve the search result.

Participant data showed that participants rated the Suman system generated keywords better than the original keywords in 63.3% of the cases and worse in 26.6% of the cases.

The Suman system is not limited by StackOverflow's 5 tags per question limit. It can add as many keywords as it can find in the document. Another advantage is that it can

add additional tags to answers. StackOverflow doesn't allow extra tags for answer. In Suman system answers are tagged separately from questions so it can have additional tags. The Suman system can also add frequency of the keywords in the questions and answers that can be used as a weight to give importance to the keywords. The other advantage is that it can add categories to the questions and answers through keywords. This creates a graph like structure and can have similar keywords associated with them.

A quick glance at the lower rated keywords shows the limitation of the Suman system. The Suman system incorporates all of the limitations of the external systems which are managed within it. One of the main drawbacks of the Suman system is that it generates the keywords and links them to the DBpedia and OpenCalais dataset. If the topic doesn't exist in the DBpedia and OpenCalais then the keywords are not linked to it and are completely ignored. This could be overcome by using an NLP library to find more keywords and not link it to any external Linked Data Cloud. It would help better categorization of the system.

The other drawback of the Suman system is that the data collected is technical data. In these datasets lots of misspelling, abbreviation and initialism are used. These colloquia are easy to understand for programmers, but the keywords generator find it difficult to interpret and link. Also, in some cases the keywords are linked to the wrong topic. Such as in one example the keyword 'Eclipse' was linked to the natural phenomenon of 'Eclipse' not the 'Eclipse (software)' topic.

The other limitation is that the versions of software and programming languages and topics like Python 2.7, Python 2.7.3, etc, is not individual page or topic, they are the section or subsection of the bigger topics. These are harder to link to the Linked Data Cloud.

One anomaly found in the keyword experiment data is that, one question was about solving a time zone problem. The main topic of the question was time zone, but the text contain the name of the cities and countries. The keyword annotating algorithm linked and annotated all the cities with its topics and gave it a high confidence rating (92%) and the main focus time zone only got 78% confidence rating. Also, one question had an image attached to it that provided additional information about the questions. And there was no way to process the image and gather the information by the keyword annotator.

Overall, the generated keywords performed better than the original keywords and provided addition information in regards to the questions but they have some limitations and drawbacks. This answered KQ3 and shows that Linked Data and Semantic Web technologies are useful and add value to the PSN dataset.

5.5.3 Q&A Correlation Test

The Suman system algorithm when searching for an answer to an unanswered question it ranks all the results and gives a score between 0 and 10. To test the usefulness of the answers, if it really does provide a solution to the unanswered question the Q&A experiment was designed.

Participants were asked to rate how good the answers provided solutions to the unanswered question between 1 and 10. The questions were of all difficulty levels and the answers were of a range of quality. If there was a positive relationship between the algorithm rating and participant rating, then it would prove that the algorithm was searching for the right answers.

For evaluating the answers an experiment was designed and all the data were collected and analyzed in SPSS. A Pearson correlation test was chosen for analysis because the data values are at regular intervals and there is a linear relationship between the two variables (algorithm's rating and participants' rating).

The Pearson correlation test was performed to measure the relationship between the participants rating and Suman algorithm rating. **There was a positive correlation between the two variables ($r = .380$, $n = 46$, p (two-tailed) = .009).**

Correlations		Users_rating	Algo_rating
Users_rating	Pearson Correlation	1	.380**
	Sig. (2-tailed)		.009
	N	46	46
Algo_rating	Pearson Correlation	.380**	1
	Sig. (2-tailed)	.009	
	N	46	46

** . Correlation is significant at the 0.01 level (2-tailed).

Figure 5.10: Answers correlation test

The correlation between the two ratings is moderately strong and the significance is $<.01$. The scatter plot diagram 5.11 summarizes the result.

The analysis shows the algorithm is quite efficient in finding the right answers. The lower left quadrant shows all the questions that got low ratings from the Suman algorithm as well as from the participants. The upper right quadrant shows all the questions

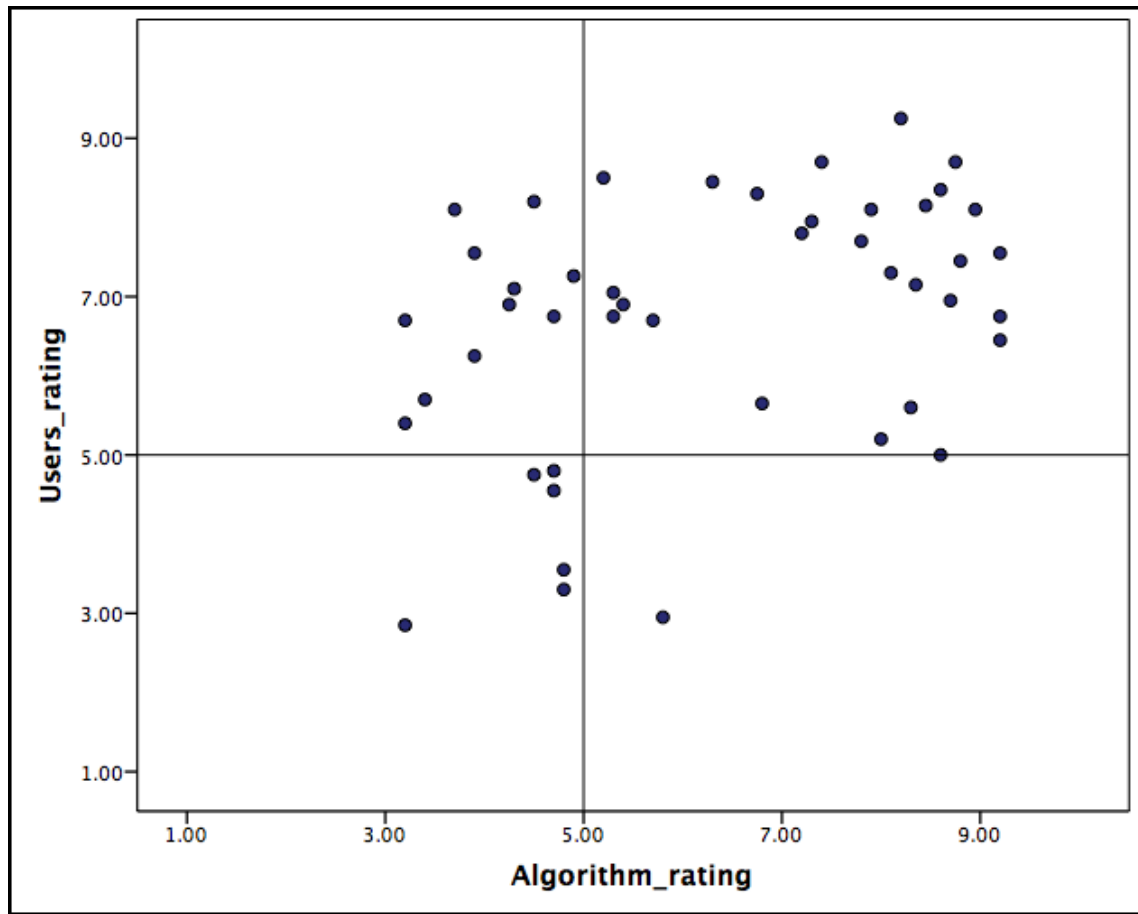


Figure 5.11: Q&A data scatter plot diagram.

that received high ratings from the Suman algorithm as well as the participants. The analysis of the data at the upper left quadrant shows that there are some answers that the participants gave high ratings but the algorithm didn't. Looking at figure 5.12 shows that they are mostly difficult questions.

There is only one question that received high ratings from the algorithm, but low rating from the participants. Looking at the question and answer it is evident that the question asked for a solution for a problem that didn't exist. The answer said so. StackOverflow users gave the high vote to that answer even though it wasn't the answer, but provided enough information to the question. Since, Suman algorithm uses crowdsourced data for the ranking, it gave that answer high confidence rating. The participant might have thought the failure of not providing an answer for the question that has no solution as a failure of the Suman system.

The participants usually agreed with the quality of the answers provided. The algorithm does well in finding the answers for difficult questions as well as easy questions as seen in the figure 5.12.

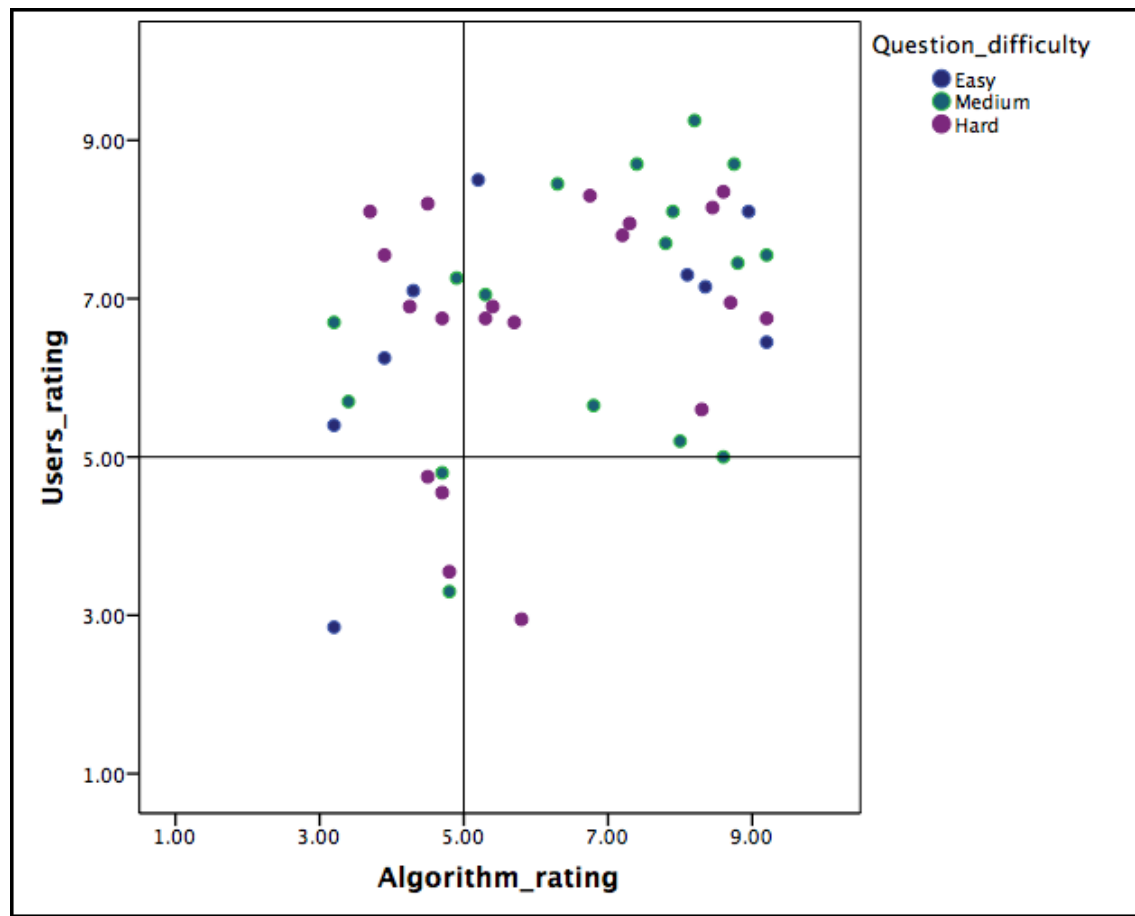


Figure 5.12: Q&A data scatter plot diagram showing questions' difficulty.

The easier questions are mostly in the upper right quadrant where participating rating is similar to the algorithm rating. It is same with the medium difficulty questions, they are mostly in the upper right quadrant.

With the questions that are hard to answer, the participants gave it a higher rating than the Suman algorithm. This could be because these questions were quite difficult and might be out of scope of the participants. They were merely guessing the answers and the answers looked valid. But there are many difficult questions in the upper right quadrant where the participants rating has correlation with the Suman algorithm's rating.

The algorithm performed well in the user evaluation, but still there are some limitations. The Suman system uses keywords and categories to first find the subset of possible answers and then performs the text search. The limitation of the keyword annotator is the limitation of the search algorithm. The algorithm doesn't perform full text search again for the remaining answers because of time and processing power restriction.

Also, the Suman system uses crowdsourced data and votes to rank the answers, so it is highly depended on people's contribution. If there are malicious users or not enough votes, then the algorithm is like a text search algorithm.

Overall, the answers performed well and provided relevant solutions to the unanswered questions. This answered KQ4 and shows that the Suman system can provide answers to unanswered questions in PSN.

5.5.4 Summary of the results

The user evaluation of the keywords showed that the Suman system generated keywords performed better than the original keywords to the system. The user evaluation of question and answer algorithm showed that participants agreed with the ranking of the algorithm, they have a positive correlation and it is moderately strong. These results help to answer the knowledge question in this thesis, specifically KQ3 and KQ4.

The Suman system has not been compared to any existing search engines and algorithms. This is because the application is built on StackOverflow and Reddit dataset. It does not have resources of popular search engines and has not been designed to be better alternative than those search engines. It is a proof of concept to show that Semantic Web and Linked Data technologies can solve problems in PSNs. The Suman algorithm combines semantic search and keyword search. The system has not been compared to popular semantic search systems because of limited resources.

The qualitative and quantitative analysis that has been done and the user evaluation showed that the Suman system performed well overall. It helped to solve the problem that the thesis focuses on.

Chapter 6

Conclusions

The Web has made it easier for people to connect and interact. It has also made it easier to study these systems. People's online activity can be studied to see how they interact and communicate and form networks.

People not only connect with their friends and families, they also interact with strangers from all over the world. They create communities with people with similar interests and expertise. They use the community to solve their problems. Forums and question-answering systems have made it easier for them to create a network, share their problems and queries and seek solutions for their problems. These systems are examples of Purposeful Social Network, which is studied in this thesis.

6.1 Summary of Research

In this thesis different types of PSNs were analyzed. People may come together with a common purpose and solve problems and create a PSN. Some PSNs are small and agile, and thrive on the user contribution. This type of system requires a strong framework to support engagement and incentives for people to contribute. These PSNs use crowdsourcing to solve problem and create a knowledge base. All these crowdsourced information can be used and additional knowledge could be added using Semantic Web and Linked Data technologies to improve the functionalities of PSNs.

6.1.1 Limitation of current systems

Many social networking services are closed. This can often be a problem for users for a number of reasons. Users may be obliged to create multiple accounts across different SNSs to use their service. There is often no easy way to merge different social networks

and their data together. Consequently, many SNS users do not have the freedom to move to different networks and take their data with them.

This problem also exists for question answering websites, forums and other PSNs where people come together and solve problems. These PSNs are crowdsourced where people create and share knowledge. In some cases questions receive no answers. The other problem faced by users is to search for answers for unanswered questions. Normal searches can be performed and different search engines can be used to look for answers to these questions, but they use text based search to find solutions and can't access data from closed systems. Also, these search engines don't use the network structure to find answers or recommend experts that can help with solving the problem.

This thesis presents a study on a programming based Q&A websites (StackOverflow) and an online forum (Reddit) as examples of PSNs. StackOverflow is a forum in which programmers ask questions about their problems and errors, and other programmers in the field answer them and provide solutions. Reddit is an online forum where users can ask questions as well as share current news and other information. These websites use crowdsourcing to find the best questions, answers and resources. Crowdsourcing is used to maintain the quality of the content, to moderate the community and to stop spam and other antisocial activities.

On StackOverflow and Reddit many questions go unanswered or do not have any comments and solutions. The people in the long tail do not get any response. These websites follow the power law of the web (Albert et al., 1999). There are few popular questions and posts that receive most of the votes and answers and the remaining posts are in the long tail that do not receive any reply or votes. The analysis done in chapter 3 shows how many posts don't get any response or votes in StackOverflow and Reddit. This is another limitation of the system addressed in this thesis.

6.1.2 Research Questions

This thesis is an investigation of whether Linked Data and Semantic Web technologies can help to answer unanswered questions on PSNs.

This research problem is broken down into smaller questions. First, PSNs are defined and studied. StackOverflow and Reddit were used as examples of PSN. This thesis studied the communication network of the users in the Reddit and StackOverflow communities, studied their motivations and incentive systems that encourage user participation. Research is also done to explore if finding answers can help the long tail of users with no answers and provide them with solutions to their problems.

This thesis also attempts to ascertain if Semantic Web technologies can be used in PSNs. These technologies can be used to integrate different PSNs by using the Semantic Web

technologies to structure the data and add semantic to it. This structured data of PSNs is linked to concepts and terms using Linked Data technologies. It is ascertained if the Semantic Web technologies can find different concepts and then categorizes the questions and answers. This could provide broader and narrower search terms. There is evidence that the broader and narrower search terms have the potential to improve accuracy in searching for answers to unanswered questions.

Similarly, this approach can also be used to find experts in the area and recommend them to get answers. These questions have been broken down into knowledge questions and design problems based on Design Science methodology and the Suman system is built to answer these questions.

6.1.3 Research Contribution

In this thesis PSNs are defined based on current literature in the field and by doing a case study on StackOverflow and Reddit, as examples of PSNs. The community structure, network ties, user motivation and current problems with these PSNs were analysed. Inference from the results of these analyses were taken into consideration to understand PSNs better and design a system to solve the existing problems mentioned above.

The Suman system was created to answer the research questions. This system searched for unanswered questions using Semantic Web and Linked Data technologies to add meaning to the PSN datasets and crowdsourced information were used to improve search and information retrieval.

Design science methodologies were followed to investigate the problems and provide a solution. The scientific steps taken to investigate, justify and evaluate the Suman system lead to design the Suman system. The websites StackOverflow and Reddit were used to collect the data. The Suman system used the APIs of the websites and collected questions, answers, posts, votes and user profiles.

The data were analyzed to study the structure of the community and a social network graph was generated. The network graphs of these websites were not formed by the explicit connection of users (friendship) but by studying user interaction and how the objects were connected with each other. The network ties, user interactions and incentive models were studied to see how a website with a small community of programmers created a self-sustaining environment for users to participate and continuously create high quality questions and answers and solve problems.

The Suman system converted the data into RDF. It cleaned the data and used ontologies (FOAF, SIOC, etc.) to structure the data and convert it into RDF.

The Suman system used Wikipedia-Miner and OpenCalais to annotate the dataset with keywords. This also addressed the name-entity disambiguation problem and helped to

classify names and entities to proper concepts and categories. These tools are based on Natural Language Processing and machine-learning techniques. They matched the keywords with Wikipedia topics and the OpenCalais vocabulary. All the keywords and topics were categorized and linked with DBpedia and the OpenCalais knowledge base. This made it possible to link the converted dataset to the Linked Data Cloud.

The Suman system also created a document keyword graph to improve search. The graph linked the keywords with all the questions, answers and experts. The links were weighted by the votes given to questions, answers and the frequency of the keywords. The data were used to improve the indexing of the dataset and SPARQL search results.

To evaluate the Suman system, unanswered questions from the websites were used as search queries and the answers were saved. The Suman system also recommended experts who were best suited to answer the question. Two experiments were designed to evaluate the keywords and the answers. In the keyword experiment, participants were shown two sets of keywords associated with the questions, the original keywords and the system generated keywords. They were asked to rate the quality of the keywords based on how well those keywords described the questions. The answer experiment showed participants an answer to an unanswered question. They were asked to rate the quality of the answers based on how well they provide solutions to the unanswered question. All the participants had experience with programming they used their own expertise to rate the keywords and answers.

The results obtained from the experiments were statistically analysed. It showed that the keywords generated by the Suman system were rated higher than the original keywords in 63.3% of the cases. The analysis also showed that the participants agreed with the algorithm rating for answers provided by the Suman system. There was a positive correlation between the two ratings ($r = .380$, $n = 46$, p (two-tailed) = .009). The correlation between the two ratings was moderately strong and the significance was $<.01$.

The Suman system showed that Semantic Web and Linked Data technologies can be used to solve the data integration problem of the current Web and can be used to integrate heterogeneous datasets. The added knowledge (semantics) and the linking of the data can help improve the search for unanswered questions in the PSNs that are overwhelmed by the number of posts and don't have answers for all the questions. Semantic Web and Linked Data technologies provide a decentralized platform where user generated knowledge can be utilized and improved. Potentially this may contribute to improving convenience for the users in these online communities. In PSNs when users post a question, the Suman system can show them similar questions with answers. Users can check those solutions to see if it satisfies their need. In the cases when the users do not receive any response to their question then the Suman system can recommend experts who are best suited to answer the question. The Suman system

expert recommendation can potentially be used to create a community for a particular purpose and solve problems based on topics and keywords. It provides a platform to integrate different forums and PSNs to improve the long tail of users that do get the help they ask from the websites.

6.2 Limitation of the Suman System

The Suman system helps to show the utility of Linked Data to solve the PSN problems but it has its limitations. The Suman system uses the StackOverflow and Reddit crowd-sourced data to get information about questions and answers quality. Any limitation of the original website data is also a limitation of our system. If the users of StackOverflow and Reddit were to stop voting or if the website were plagued by spammers then the Suman system would not be able to mitigate the situation.

The Suman system uses external tools like Wikipedia Miner and OpenCalais to annotate the text and find the best match for DBpedia and OpenCalais topics to link. Any limitation of these tools is a limitation of the Suman System. Even though the system only accepts keywords and links that have a confidence score greater than 50%, sometimes the keywords are linked to the incorrect concepts. If there is no Wikipedia or OpenCalais article for a concept, then those keywords are ignored by the system, even though they are valid keywords.

Currently, the system does not have a user interface. It uses the websites unanswered questions as search terms. Currently, users cannot ask their own questions to search for similar answers and questions. The Suman system runs on a terminal. There are no user interface for people to rate the quality of the search result. These limitations could potentially be improved in the future by adding a top layer GUI over the search algorithm that provides a portal for user input and feedback to improve the accuracy of the search results.

Participants in the Suman system evaluation experiment did not evaluate the Expert Finder application. There was no easy and simple way to provide a complete user profile of every user to the participants in the experiment. The user profile would consist of user's posts, badges, votes and reputation points. This is a huge data and there was no easy way to show it to the participants and then ask them to rate the expert recommender. The same algorithm that is used to improve the search of answers is used to search for experts.

Search engines such as Google and Bing use sophisticated text search and algorithms such as PageRank. They crawl the entire web and find the best match for the search result. Due to limited resources, the Suman system only uses two websites and provides answers from those datasets. This is a proof of concept. The results do not necessarily

indicate that the Suman system provides a better result than the current search engines. The results suggest that the Semantic Web and Linked Data technologies provide useful tools for data integration. These have potential to improve the search results as more knowledge bases and communities are linked together.

6.3 Future Work

The Suman system currently does not have any user interface. A user interface would let people enter their own questions and help in finding similar questions and answers from StackOverflow and Reddit. Another layer of crowdsourcing on top of the system could potentially allow users to up-vote and down-vote answers and improve the search result.

Currently, the Suman system recommends experts who are best suited to answer a question. The system could be extended to create community and an agile PSN. People could use the system to find the right experts to solve their problem, get details of the experts and contact them. There is scope for improving the system by adding more communities and website data to the knowledge base. People can create an agile PSN and find people in different communities and thus potentially extend their own social network.

Currently, the Suman system uses external tools for data annotation to find important topics and categories and link it to the Linked Data cloud. Ongoing research in Natural Language Processing may help in finding deeper semantics, synonyms and categories to the text. This layer could be on top of the annotation layer. The extracted keywords and extra information from the deep learning algorithms might not be linked to the Linked Data Cloud, but it could potentially improve the document keyword graph. This might improve the overall search results by finding better questions and answers in broader and narrower topics.

For future work, the Suman System could potentially be improved by adding more datasets and linking to different knowledge bases. The Suman System in its current form provides a tool for data integration across platforms. It is possible to create a unified knowledge base from different networks on the Web.

Some of the data on the Web are freely available, but most of the data is still bound behind the closed walls of different websites. Using the APIs of these websites and with the participation of users, it is possible to free user data from the ubiquitous SNS silos. The Solid project is a step in this direction (Conner-Simons, 2015). These different datasets can be integrated using Semantic Web and Linked Data technologies. The findings of this thesis suggest that this may open up opportunities to improve the community experience on PSNs in the future.

Appendix A

Experiment Questionnaire

The Suman system was evaluated by doing two experiments. The first keywords experiment asked participants to rate original keywords and system generated keywords based on how well they describe the questions. The second answers experiment asked participants to rate answers based on how well they answer the unanswered questions.

The questions used in these two experiments are mentioned below.

A.1 Keywords Experiment

Keywords experiments showed participants 30 questions. Each question had original keywords and the Suman system generated keywords. Users were asked to rate both sets of keywords in the scale of 1 to 5. 15 questions were asked from programming language Python¹ and another 15 were from programming language Java². All the 30 questions with instructions are below.

Instruction for question:

1. Read the question (section A)
2. Rate how well the keywords describe the question in your opinion (Section B)

1. A. Read text from PNG with standard lib

Is there a way to read text from a PNG-File in Python by using only the standard libraries Python provides?

¹<https://www.isurvey.soton.ac.uk/start.php?id=13709>

²<https://www.isurvey.soton.ac.uk/start.php?id=57376>

B. Keywords: python, png(portable network graphics), programming language, library (computing)

Q. Please rate how well the keywords describe the question in your opinion (Section B)

Very bad ☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 **Very good**

2. **A. Read text from PNG with standard lib**

Is there a way to read text from a PNG-File in Python by using only the standard libraries Python provides?

B. Keywords: python

Q. Please rate how well the keywords describe the question in your opinion (Section B)

Very bad ☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 **Very good**

3. **A. How to programmatically create dummy email thread?**

For comparing some software I need a thread of mails, i.e. some mails with replies and replies to the replies... Content does not matter, but attachment and richt-text would be nice.

I wonder how to create such a dummy mail thread programmatically (preferably using Linux commandline tools or Python).

How would I create those dummy mails?

B. Keywords: python, linux, email, cli, dummy-data

Q. Please rate how well the keywords describe the question in your opinion (Section B)

Very bad ☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 **Very good**

4. **A. How to programmatically create dummy email thread?**

For comparing some software I need a thread of mails, i.e. some mails with replies and replies to the replies... Content does not matter, but attachment and richt-text would be nice.

I wonder how to create such a dummy mail thread programmatically (preferably using Linux commandline tools or Python).

How would I create those dummy mails?

B. Keywords: python, programming language, email, computer software, linux, command-line interface, dummy, mail

Q. Please rate how well the keywords describe the question in your opinion (Section B)

Very bad ☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 **Very good**

5. A. How to call python code from excel?

I have written a python code using selenium for automating files extraction from website and then I need to format these reports and append them and i am using macros for this.

Is there any way to call the python code from excel by integrating it with VBA or something like that?

B. Keywords: python (programming language), vba (visual basic for applications), selenium, macro (computer science), code, automation, website

Q. Please rate how well the keywords describe the question in your opinion (Section B)

Very bad ☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 **Very good**

6. A. How to call python code from excel?

I have written a python code using selenium for automating files extraction from website and then I need to format these reports and append them and i am using macros for this.

Is there any way to call the python code from excel by integrating it with VBA or something like that?

B. Keywords: excel, excel-vba, python-2.7

Q. Please rate how well the keywords describe the question in your opinion (Section B)

Very bad ☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 **Very good**

7. A. In Python IDLE show completion what does -i mean?

I managed to get Show Completion to work thanks to this answer.
 But what does `str(object) -> string` mean as a tip after typing the opening bracket
 Example code:

```
linkText = "some text"
elms = browser.find_elements(By.PARTIAL_LINK_TEXT(linkText))
On Run gives: TypeError: 'str' object is not callable
```

Does it mean `linkText` should be a pointer to string? How do I enter a pointer in Python?

B. Keywords: python, python-idle

Q. Please rate how well the keywords describe the question in your opinion (Section B)

Very bad ☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 Very good

8. **A. In Python IDLE show completion what does `-i` mean?** I managed to get Show Completion to work thanks to this answer.

But what does `str(object) -i string` mean as a tip after typing the opening bracket of a function?

Example code:

```
linkText = "some text"
elms = browser.find_elements(By.PARTIAL_LINK_TEXT(linkText))
On Run gives: TypeError: 'str' object is not callable
```

Does it mean `linkText` should be a pointer to string? How do I enter a pointer in Python?

B. Keywords: python, string (computer science), pointer (computing), object (computer science)

Q. Please rate how well the keywords describe the question in your opinion (Section B)

Very bad ☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 Very good

9. **A. python significant figures and float**

I'm do not understand why $1.1 + 2.2$ is not 3.3 if a computer calculates this. I am trying to understand the working of binary floating points.. but I am not even sure of that float the cause is. could you explain this to me?, I have not been able to find a clear explanation.

```
Python 2.7.4 (default, Apr 6 2013, 19:54:46) [MSC v.1500 32 bit (Intel)] on
Type "copyright", "credits" or "license()" for more information.
>>> ===== RESTART =====
>>>
>>> 1.1+2.2
3.3000000000000003
>>>
```

B. Keywords: python, floating-point, significant-digits

Q. Please rate how well the keywords describe the question in your opinion (Section B)

Very bad ☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 **Very good**

10. **A. python significant figures and float**

I'm do not understand why $1.1 + 2.2$ is not 3.3 if a computer calculates this. I am trying to understand the working of binary floating points.. but I am not even sure of that float the cause is. could you explain this to me?, I have not been able to find a clear explanation.

```
Python 2.7.4 (default, Apr 6 2013, 19:54:46) [MSC v.1500 32 bit (Intel)] on
Type "copyright", "credits" or "license()" for more information.
>>> ===== RESTART =====
>>>
>>> 1.1+2.2
3.3000000000000003
>>>
```

B. Keywords: python(programming language), significant figures, float, bit, intel, windows api, computer, copyright

Q. Please rate how well the keywords describe the question in your opinion (Section B)

Very bad ☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 Very good

11. **A. How to verify ssl certificate?**

im working on dns masking. Im asking user to provide URL for masking and SSL certificate. I need to validate whether the certificate user provided is of valid format or not. The problem is if user doesnt provide valid certificate apache doesn't restart. Im using python.

I've tried some solutions provided here but couldn't validate ssl certificate.

Any help will be highly appreciated.

Thanks!

B. Keywords: python, tls (transport layer security), url (uniform resource locator), dns (domain name system), apache

Q. Please rate how well the keywords describe the question in your opinion (Section B)

Very bad ☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 Very good

12. **A. How to verify ssl certificate?**

im working on dns masking. Im asking user to provide URL for masking and SSL certificate. I need to validate whether the certificate user provided is of valid format or not. The problem is if user doesnt provide valid certificate apache doesn't restart. Im using python.

I've tried some solutions provided here but couldn't validate ssl certificate.

Any help will be highly appreciated.

Thanks!

B. Keywords: python, validation, ssl-certificate

Q. Please rate how well the keywords describe the question in your opinion (Section B)

Very bad ☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 Very good

13. **A. HTTP Request coming up blank**

With my first foray into anything-related-to-web programming, I am using the Requests external library (Requests), I am testing it on <https://google.com.au> to see if I'm doing it right.

```
import requests

proxy_dict = {
    "https" : "10.10.20.99:8080"
}

r = requests.get('https://google.com.au', proxies = proxy_dict)
print r.content
print r.status_code
with output

<HTML>>
200
```

i.e. a completed HTTP request yet with no returned HTML information. I have read the relevant docs for "Requests" but I can't get the same results as that example (which is the same as this one, except using github.com instead of google.com.au). I am a complete and utter noob at all things HTTP/HTML right now, so does anyone have some idea where I'm going wrong?

Thanks!

EDIT: I forgot to mention, this will be behind some sort of company firewall as I am doing this through work. That proxy is my work's proxy.!

B. Keywords: python, html, http, newbie

Q. Please rate how well the keywords describe the question in your opinion (Section B)

Very bad ☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 **Very good**

14. **A. HTTP Request coming up blank**

With my first foray into anything-related-to-web programming, I am using the Requests external library (Requests), I am testing it on https://google.com.au to see if I'm doing it right.

```
import requests

proxy_dict = {
    "https" : "10.10.20.99:8080"
}
```

```

r = requests.get('https://google.com.au', proxies = proxy_dict)
print r.content
print r.status_code
with output

<HTML>>
200

```

i.e. a completed HTTP request yet with no returned HTML information. I have read the relevant docs for "Requests" but I can't get the same results as that example (which is the same as this one, except using github.com instead of google.com.au). I am a complete and utter noob at all things HTTP/HTML right now, so does anyone have some idea where I'm going wrong?

Thanks!

EDIT: I forgot to mention, this will be behind some sort of company firewall as I am doing this through work. That proxy is my work's proxy.!

B. Keywords: python

Q. Please rate how well the keywords describe the question in your opinion (Section B)

Very bad ☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 **Very good**

15. **A. Python Java Integration**

I'm developing a program completely written in Python but I need to integrate a Java code inside the file console.py. I want to integrate Sphinx4's program to give GNS3 the capability of voice recognition.

Is it possible? What do I need to do this!

B. Keywords: java, python, voice-recognition, jpype

Q. Please rate how well the keywords describe the question in your opinion (Section B)

Very bad ☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 **Very good**

16. **A. Python Java Integration**

I'm developing a program completely written in Python but I need to integrate a Java code inside the file console.py. I want to integrate Sphinx4's program to give GNS3 the capability of voice recognition.

Is it possible? What do I need to do this!

B. Keywords: python (programming language), java (programming language), speech recognition

Q. Please rate how well the keywords describe the question in your opinion (Section B)

Very bad ☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 **Very good**

17. **A. What is the most painless way to remove one version of python and not the other on OSX 10.7.4?**

I accidentally installed 64 bit python a few weeks ago, and now all of my programs jump for it by default, instead of the more functional 32 bit python. Is there a painless way to uninstall 64 bit python but not 32?

Thanks!

B. Keywords: python, mac os x, uninstaller, bit

Q. Please rate how well the keywords describe the question in your opinion (Section B)

Very bad ☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 **Very good**

18. **A. What is the most painless way to remove one version of python and not the other on OSX 10.7.4?**

I accidentally installed 64 bit python a few weeks ago, and now all of my programs jump for it by default, instead of the more functional 32 bit python. Is there a painless way to uninstall 64 bit python but not 32?

Thanks!

B. Keywords: python, osx, uninstall

Q. Please rate how well the keywords describe the question in your opinion (Section B)

Very bad ☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 Very good

19. **A. Free memory in python run time**

I have a python server running on tornado , there is request that uses lots of memory (using a gdata library) . I deleted those object after using them and even did `gc.collect()` , But when i see the system memory using `free -m` , I could see memory increase in memory when i do those operations using gdata . But the memory does not get freed up , after i delete the object . The memory gets freed up only after the main python program is killed . I want to know if there is any way to free up the memory .

B. Keywords: python

Q. Please rate how well the keywords describe the question in your opinion (Section B)

Very bad ☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 Very good

20. **A. Free memory in python run time**

I have a python server running on tornado , there is request that uses lots of memory (using a gdata library) . I deleted those object after using them and even did `gc.collect()` , But when i see the system memory using `free -m` , I could see memory increase in memory when i do those operations using gdata . But the memory does not get freed up , after i delete the object . The memory gets freed up only after the main python program is killed . I want to know if there is any way to free up the memory .

B. Keywords: python, ram, run time (program lifecycle phase), server (coumputing), library (computing)

Q. Please rate how well the keywords describe the question in your opinion (Section B)

Very bad ☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 Very good

21. **A. URL links in Eclipse python code**

I like to document my python code with links to web pages that explain the algorithm in detail. But I can't make the links 'live', so that just clicking on them in Eclipse brings up the page in the platform browser.

I've just been putting the URL's inside the 3-double-quote pair, then having to manually copy paste them to the browser.

Is there a better way?

B. Keywords: python, eclipse (software), url, web browser, algorithm, pydev

Q. Please rate how well the keywords describe the question in your opinion (Section B)

Very bad ☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 **Very good**

22. **A. URL links in Eclipse python code**

I like to document my python code with links to web pages that explain the algorithm in detail. But I can't make the links 'live', so that just clicking on them in Eclipse brings up the page in the platform browser.

I've just been putting the URL's inside the 3-double-quote pair, then having to manually copy paste them to the browser.

Is there a better way?

B. Keywords: python, eclipse, pydev

Q. Please rate how well the keywords describe the question in your opinion (Section B)

Very bad ☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 **Very good**

23. **A. need a time widget or clock widget**

is there anything like a timewidget or clock widget in Django which will help me to input data into a form for appointment date (like we have calendar or date widget).

i have my date widget working with the below code in my forms.py:

```
import datetime
from django.forms.extras.widgets import SelectDateWidget
mydate = forms.DateField(widget=SelectDateWidget)
```

is there anything like this for time widget

B. Keywords: python, django (web framework), gui widget, clock, calendar, data, import

Q. Please rate how well the keywords describe the question in your opinion (Section B)

Very bad ☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 **Very good**

24. A. need a time widget or clock widget

is there anything like a timewidget or clock widget in Django which will help me to input data into a form for appointment date (like we have calendar or date widget).

i have my date widget working with the below code in my forms.py:

```
import datetime
from django.forms.extras.widgets import SelectDateWidget
mydate = forms.DateField(widget=SelectDateWidget)
```

is there anything like this for time widget

B. Keywords: python, django, time, django-widget

Q. Please rate how well the keywords describe the question in your opinion (Section B)

Very bad ☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 **Very good**

25. A. Python child process limits

I'd like to create a process in Python (probably with subprocess and Popen), which should have limited CPU time, limited child processess and memory bandwidth. I can;t find a way to do this. resource.setrlimit does not seem to work.

My code is :

```
import os
import sys
import resource
import subprocess
import signal

def setlimits():
    os.seteuid(65534) # Has to run as root user in order to be able to setuid
    resource.setrlimit(resource.RLIMIT_CPU, (1, 1))
```

```
resource.setrlimit(resource.RLIMIT_FSIZE, (500, 500))
resource.setrlimit(resource.RLIMIT_NPROC, (80, 80))
```

```
p = subprocess.Popen( ["./exec.out"] , preexec_fn=setlimits )
```

B. Keywords: python, process (computing), cpu, child process, exec (operating system), operating system, bandwidth, setuid

Q. Please rate how well the keywords describe the question in your opinion (Section B)

Very bad ☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 **Very good**

26. **A. Python child process limits**

I'd like to create a process in Python (probably with subprocess and Popen), which should have limited CPU time, limited child processes and memory bandwidth. I can;t find a way to do this. resource.setrlimit does not seem to work.

My code is :

```
import os
import sys
import resource
import subprocess
import signal

def setlimits():
    os.seteuid(65534) # Has to run as root user in order to be able to setuid
    resource.setrlimit(resource.RLIMIT_CPU, (1, 1))
    resource.setrlimit(resource.RLIMIT_FSIZE, (500, 500))
    resource.setrlimit(resource.RLIMIT_NPROC, (80, 80))

p = subprocess.Popen( ["./exec.out"] , preexec_fn=setlimits )
```

B. Keywords: python, process, fork, limits

Q. Please rate how well the keywords describe the question in your opinion (Section B)

Very bad ☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 **Very good**

27. A.Using C++ library functions from Python

I got a library created for Win x86 with MS Visual Studio 2010. And I can not change the content of a library to use Boost.Python.

I'm using Python 3.3 with PyQt4 to create interface, but not restricted to these versions. I need to call functions and get objects from said C++ library. What is the easiest way to wrap C++ library to be called from python?

I guess, that such question was already asked, but I can not seem to find it.

Here's an example of header file:

```
namespace SDK
{
    class IMethod
    {
    public:
        virtual IModel* CreateModel(const IBuffer* pBuffer, const char* text) = 0;
    };

    extern __declspec(dllexport) SDK::IMethod* CreateMethod(MethodID integer);
}
```

B. Keywords: python, c++, windows, dll, pyqt4

Q. Please rate how well the keywords describe the question in your opinion (Section B)

Very bad ☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 **Very good**

28. A.Using C++ library functions from Python

I got a library created for Win x86 with MS Visual Studio 2010. And I can not change the content of a library to use Boost.Python.

I'm using Python 3.3 with PyQt4 to create interface, but not restricted to these versions. I need to call functions and get objects from said C++ library. What is the easiest way to wrap C++ library to be called from python?

I guess, that such question was already asked, but I can not seem to find it.

Here's an example of header file:

```
namespace SDK
{
    class IMethod
```

```

    {
    public:
        virtual IModel* CreateModel(const IBuffer* pBuffer, const char* text)
    };

    extern __declspec(dllexport) SDK::IMethod* CreateMethod(MethodID integer)
}

```

B. Keywords: python, c++, pyqt, dynamic-link library (dll), class (computer programming), microsoft visual studio, sdk, header file

Q. Please rate how well the keywords describe the question in your opinion (Section B)

Very bad ☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 **Very good**

29. **A. python pprint: how to output utf8 characters?**

I have a dictionary with 'utf8 char' : number . pprint() would treat the utf8 as a byte array and output hex values. Is it possible to tell it to print it as string so that the console can render UTF8 text?

B. Keywords: python, string(computer science), character (computing), dictionary, hexadecimal, utf-8

Q. Please rate how well the keywords describe the question in your opinion (Section B)

Very bad ☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 **Very good**

30. **A. python pprint: how to output utf8 characters?**

I have a dictionary with 'utf8 char' : number . pprint() would treat the utf8 as a byte array and output hex values. Is it possible to tell it to print it as string so that the console can render UTF8 text?

B. Keywords: python

Q. Please rate how well the keywords describe the question in your opinion (Section B)

Very bad ☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 **Very good**

31. A. Java, console.readPassword adds extra line. How to delete it?

When i use console.readPassword() to read user passwords through console, there is always one line added to the console.

How to disable this behavior or how to delete that extra line (and move the cursor after the last character in the line before)? What escape character to use?

Thanks

B. Keywords: java, console, password

Q. Please rate how well the keywords describe the question in your opinion (Section B)

Very bad ☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 **Very good**

32. A. Java, console.readPassword adds extra line. How to delete it?

When i use console.readPassword() to read user passwords through console, there is always one line added to the console.

How to disable this behavior or how to delete that extra line (and move the cursor after the last character in the line before)? What escape character to use?

Thanks

B. Keywords: java, object oriented programming language, console, password, cursor (computers), escape character

Q. Please rate how well the keywords describe the question in your opinion (Section B)

Very bad ☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 **Very good**

33. A. is there an equivalent of Python's timeit module in Java stdlib

I was amazed at Python's timeit module and wondered if there's an equivalent in Java's standard library. If not, is there a 3rd party module?

B. Keywords: java, python, programming language, library (computing)

Q. Please rate how well the keywords describe the question in your opinion (Section B)

Very bad ☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 **Very good**

34. **A. is there an equivalent of Python's timeit module in Java stdlib**

I was amazed at Python's timeit module and wondered if there's an equivalent in Java's standard library. If not, is there a 3rd party module?

B. Keywords: java, python

Q. Please rate how well the keywords describe the question in your opinion (Section B)

Very bad ☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 **Very good**

35. **A. java.lang.NoSuchMethodError when trying to read .xslm file**

Exception in thread "main" java.lang.NoSuchMethodError: org.apache.xmlbeans.XmlOptions.set
at org.apache.poi.POIXMLDocumentPart.<clinit>(POIXMLDocumentPart.java:56)
at rulebooksToExcel.GenerateExcel.generateExcel(GenerateExcel.java:34)
at rulebooksToExcel.ParseNortDocFiles.main(ParseNortDocFiles.java:165)

I am getting the error at :

workbook = new XSSFWorkbook(in); I read other similar questions but they all suggest XMLBeans Version 2.0+. But I am using 2.6, and I can't find any other explanation for what might be causing this.

B. Keywords: java, apache, thread (coumputing), XMLBeans

Q. Please rate how well the keywords describe the question in your opinion (Section B)

Very bad ☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 **Very good**

36. **A. java.lang.NoSuchMethodError when trying to read .xslm file**

Exception in thread "main" java.lang.NoSuchMethodError: org.apache.xmlbeans.XmlOptions.set
at org.apache.poi.POIXMLDocumentPart.<clinit>(POIXMLDocumentPart.java:56)
at rulebooksToExcel.GenerateExcel.generateExcel(GenerateExcel.java:34)
at rulebooksToExcel.ParseNortDocFiles.main(ParseNortDocFiles.java:165)

I am getting the error at :

workbook = new XSSFWorkbook(in); I read other similar questions but they all suggest XMLBeans Version 2.0+. But I am using 2.6, and I can't find any other explanation for what might be causing this.

B. Keywords: java, apache-poi, xmlbeans, xslm

Q. Please rate how well the keywords describe the question in your opinion (Section B)

Very bad ☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 Very good

37. **A. How do we call a method in a jar file from a java script method**

How do we call a method in a jar file from a java script method. I am using a third party to authenticate a HTML5 application, the jquery method redirects to the third party url and they redirect back to the app after validation at their end. Now in Java we can use the methods in their jar file to get back the User ID, but I am not sure if we can do it using js.

Code in java, the jar is added in the classpath -

```
private static String UID(HttpServletRequest req) {
    String unEncCookie = null;
    String cookie = getSecCookie(req);
    if (cookie == null)
        return null;
    else {
        unEncCookie = JAR.JAR(cookie, "param1", "param2");
        if (unEncCookie == null || "".equals(unEncCookie))
            return null;
        else{
            return unEncCookie.split("|")[5]; // 6th value is UID
        }
    }
}
```

B. Keywords: java, javascript, jar file, HTML5, authentication, URL, data validation, string, HTTP cookie

Q. Please rate how well the keywords describe the question in your opinion (Section B)

Very bad ☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 Very good

38. **A. How do we call a method in a jar file from a java script method**

How do we call a method in a jar file from a java script method. I am using a third party to authenticate a HTML5 application, the jquery method redirects to the third party url and they redirect back to the app after validation at their end. Now in Java we can use the methods in their jar file to get back the User ID, but I am not sure if we can do it using js.

Code in java, the jar is added in the classpath -

```
private static String UID(HttpServletRequest req) {
    String unEncCookie = null;
    String cookie = getSecCookie(req);
    if (cookie == null)
        return null;
    else {
        unEncCookie = JAR.JAR(cookie, "param1", "param2");
        if (unEncCookie == null || "".equals(unEncCookie))
            return null;
        else{
            return unEncCookie.split("|")[5]; // 6th value is UID
        }
    }
}
```

B. Keywords: java, javascript, jar

Q. Please rate how well the keywords describe the question in your opinion (Section B)

Very bad ☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 Very good

39. **A. Is there any Java API for PCRE regex maching**

we have a requirement to match the PCRE expression against a file name. Tried to find the API regarding it, but unable to find any. Need some suggestion how to achieve it. If any of you have already done, a sample would be more grateful

B. Keywords: java, pcre

Q. Please rate how well the keywords describe the question in your opinion (Section B)

Very bad ☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 Very good

40. **A. Is there any Java API for PCRE regex matching**

we have a requirement to match the PCRE expression against a file name. Tried to find the API regarding it, but unable to find any. Need some suggestion how to achieve it. If any of you have already done, a sample would be more grateful

B. Keywords: java, programming language, perl compatible regular expressions (pcre), regular expression, api, list of java APIs

Q. Please rate how well the keywords describe the question in your opinion (Section B)

Very bad ☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 Very good

41. **A. Play Framework 2.1.3 Debugger failed to attach: handshake failed**

I get this message in the console when I run a Play.

```
C:\Users\Daniel\desenvolvimento\Sistemas>play debug
Listening for transport dt_socket at address: 9999
Debugger failed to attach: handshake failed - received >POST /setwindo<-
expected >JDWP-Handshake<
Debugger failed to attach: handshake failed - received >POST /setwindo<-
expected >JDWP-Handshake<
```

I don't have any problems running the app but this "Debugger failed" message keeps coming out and it just bugs me.

This happens even if I create a clean project.

Play 2.1.3 Windows7 64bit

How could I get rid of this message?

B. Keywords: java, playframework, playframework-2.1

Q. Please rate how well the keywords describe the question in your opinion (Section B)

Very bad ☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 Very good

42. **A. Play Framework 2.1.3 Debugger failed to attach: handshake failed**

I get this message in the console when I run a Play.

```
C:\Users\Daniel\desenvolvimento\Sistemas>play debug
Listening for transport dt_socket at address: 9999
Debugger failed to attach: handshake failed - received >POST /setwindo<-
expected >JDWP-Handshake<
Debugger failed to attach: handshake failed - received >POST /setwindo<-
expected >JDWP-Handshake<
```

I don't have any problems running the app but this "Debugger failed" message keeps coming out and it just bugs me.

This happens even if I create a clean project.

Play 2.1.3 Windows7 64bit

How could I get rid of this message?

B. Keywords: java, windows 7, play framework, debugger, application software

Q. Please rate how well the keywords describe the question in your opinion (Section B)

Very bad ☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 **Very good**

43. **A. Retrieve a list of all packages in a project with Google reflections library?**

I'm talking about the reflections lib. Is there any possibility to get a list of all packages which are included in the project where I let the code compile?

I've tried it with the following code bracket but I don't want to insert a project name.

```
Reflections reflections = new Reflections(/* Project name here... */);
Set<Class extends Object>> allClasses =
    reflections.getSubTypesOf(Object.class);
```

B. Keywords: java, packages, google-reflections

Q. Please rate how well the keywords describe the question in your opinion (Section B)

Very bad ☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 **Very good**

44. **A. Retrieve a list of all packages in a project with Google reflections library?**

I'm talking about the reflections lib. Is there any possibility to get a list of all packages which are included in the project where I let the code compile?

I've tried it with the following code bracket but I don't want to insert a project name.

```
Reflections reflections = new Reflections(/* Project name here... */);
Set<Class extends Object>> allClasses =
    reflections.getSubTypesOf(Object.class);
```

B. Keywords: java, programming language, google, library (computing), java package

Q. Please rate how well the keywords describe the question in your opinion (Section B)

Very bad ☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 **Very good**

45. **A. Choose Gradle JDK in IntelliJ**

My system JAVA_HOME and PATH are pointing to JDK 7, but I want to use JDK 8 in project opened in IntelliJ. I changes settings in Project Structure and it works great in IDE, but unfortunately Gradle build run from IDE still uses JDK 7. How can I specify Gradle JDK in IntelliJ 13.0?

B. Keywords: java, java development kit (JDK), IntelliJ IDEA, gradle, integrated development environment (IDE)

Q. Please rate how well the keywords describe the question in your opinion (Section B)

Very bad ☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 **Very good**

46. **A. Choose Gradle JDK in IntelliJ**

My system JAVA_HOME and PATH are pointing to JDK 7, but I want to use JDK 8 in project opened in IntelliJ. I changes settings in Project Structure and it works great in IDE, but unfortunately Gradle build run from IDE still uses JDK 7. How can I specify Gradle JDK in IntelliJ 13.0?

B. Keywords: java, intellij-idea, gradle

Q. Please rate how well the keywords describe the question in your opinion (Section B)

Very bad ☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 **Very good**

47. **A. Serve Images from a separate location in JSF1.1**

We are having a lot of images which all exist in our webroot. Is there a setting in JSF 1.1 which allows us to set location of image path.

B. Keywords: java, jsf

Q. Please rate how well the keywords describe the question in your opinion (Section B)

Very bad ☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 **Very good**

48. **A. Serve Images from a separate location in JSF1.1**

We are having a lot of images which all exist in our webroot. Is there a setting in JSF 1.1 which allows us to set location of image path.

B. Keywords: java, java server faces (jsf), programming language

Q. Please rate how well the keywords describe the question in your opinion (Section B)

Very bad ☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 **Very good**

49. **A. Avoid direct links to login page**

I'm using spring security including remember-me feature:

Now when i check remember-me checkbox, i can close browser and go to some app pages skipping the login page.

But i still can type myApp/login.jsp and go to the login page.

Is there any ways to avoid it? - I want to avoid any direct links to login page. User should be able to see login screen only if he hasn't logged in still, hasn't pressed "remember-me" button (and closed browser) or pressed logout button of my app.

B. Keywords: java, java server pages (JSP), spring framework, login, security, web browser

Q. Please rate how well the keywords describe the question in your opinion (Section B)

Very bad ☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 **Very good**

50. **A. Avoid direct links to login page**

I'm using spring security including remember-me feature:

Now when i check remember-me checkbox, i can close browser and go to some app pages skipping the login page.

But i still can type myApp/login.jsp and go to the login page.

Is there any ways to avoid it? - I want to avoid any direct links to login page. User should be able to see login screen only if he hasn't logged in still, hasn't pressed "remember-me" button (and closed browser) or pressed logout button of my app.

B. Keywords: java, jsp, spring-security, remember-me

Q. Please rate how well the keywords describe the question in your opinion (Section B)

Very bad ☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 **Very good**

51. **A. How can I convert timezone of one City to another?**

I want to convert timezone of one city to another, using Joda DateTime, the process is straight forward, but I have a problem.

I am referring to Joda Database and I am unable to locate specific cities, in the database there are so few mappings say for e.g. Asia/Kolkata, if I want to search timezone for say another city like Asia/New Delhi which is in Asia or more specifically India, I cannot do that, it is mostly probable that cities within the same country have same time zone but its not always true, so I am stuck in understanding of as how to map a city to its country and vice versa.

But there is no mapping found for cities other than what is speicified in the database.

Is there any way other than or including Joda Datetime which I can use to convert by directly inputting the city name in Olson format?

I have seen many websites 1, 2 which do what I want, please help me understand what I am missing here.

B. Keywords: java, datetime, time, timezone, jodatetime

Q. Please rate how well the keywords describe the question in your opinion (Section B)

Very bad ☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 **Very good**

52. **A. How can I convert timezone of one City to another?**

I want to convert timezone of one city to another, using Joda DateTime, the process is straight forward, but I have a problem.

I am referring to Joda Database and I am unable to locate specific cities, in the database there are so few mappings say for e.g. Asia/Kolkata, if I want to search timezone for say another city like Asia/New Delhi which is in Asia or more specifically India, I cannot do that, it is mostly probable that cities within the same country have same time zone but its not always true, so I am stuck in understanding of as how to map a city to its country and vice versa.

But there is no mapping found for cities other than what is speicified in the database.

Is there any way other than or including Joda Datetime which I can use to convert by directly inputting the city name in Olson format?

I have seen many websites 1, 2 which do what I want, please help me understand what I am missing here.

B. Keywords: java, time zone, map, database, asia, website

Q. Please rate how well the keywords describe the question in your opinion (Section B)

Very bad ☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 **Very good**

53. **A. install java plugin inside HTML element in ActionScript3**

i have tried to open a link inside HTML element in AS3 , but there is nothing . in chrome,mozilla will work and open the link because it need a java plugin installed. can i add java plugin in my Adobe Air to use it in HTML element and open link ? the link is Click Here

B. Keywords: java, html, action script, adobe integrated runtime, html element, plug-in (computing), adobe systems

Q. Please rate how well the keywords describe the question in your opinion (Section B)

Very bad ☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 **Very good**

54. **A. install java plugin inside HTML element in ActionScript3**

i have tried to open a link inside HTML element in AS3 , but there is nothing . in chrome,mozilla will work and open the link because it need a java plugin installed. can i add java plugin in my Adobe Air to use it in HTML element and open link ? the link is [Click Here](#)

B. Keywords: java, html, actionscript-3, plugins

Q. Please rate how well the keywords describe the question in your opinion (Section B)

Very bad ☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 **Very good**

55. **A. Is it possible to configure a custom error message per converter?**

Is it possible to configure a custom error message per converter?

For example I have this in my xwork-conversion.properties:

```
java.util.Date=mx.com.afirme.midas2.converter.DateConverter
```

Whenever a Date conversion fails in any action I'd like to show a message like this:

Incorrect format, expected mm/dd/yyyy

I don't want to define a custom message per property as mentioned in the documentation:

However, sometimes you may wish to override this message on a per-field basis. You can do this by adding an `i18n` key associated with just your action (Action.properties) using the pattern `invalid.fieldvalue.xxx`, where `xxx` is the field name.

B. Keywords: java, struts2

Q. Please rate how well the keywords describe the question in your opinion (Section B)

Very bad ☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 Very good

56. **A. Is it possible to configure a custom error message per converter?**

Is it possible to configure a custom error message per converter?

For example I have this in my xwork-conversion.properties:

java.util.Date=mx.com.afirme.midas2.converter.DateConverter Whenever a Date conversion fails in any action I'd like to show a message like this:

Incorrect format, expected mm/dd/yyyy

I don't want to define a custom message per property as mentioned in the documentation:

However, sometimes you may wish to override this message on a per-field basis. You can do this by adding an i18n key associated with just your action (Action.properties) using the pattern invalid.fieldvalue.xxx, where xxx is the field name.

B. Keywords: Java, internationalization and localization (i18n), dd (unix), error message

Q. Please rate how well the keywords describe the question in your opinion (Section B)

Very bad ☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 Very good

57. **A. JTable Customized Row**

I would like some input on how to implement the following row rendering. I don't need actual code but just the concept. Row 1 is Yellow and Row 2 is White. My bean has Order Date, Req Ship Date, ... and Product as properties.

Order Date	Req Ship Date	InHand Date	Customer	Ship Date	WO#	Ship By	Ready	Shipped	Notes
Product:									
Product:									

B. Keywords: java, swing, table, tablecellrenderer

Q. Please rate how well the keywords describe the question in your opinion (Section B)

Very bad ☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 Very good

58. **A. JTable Customized Row**

I would like some input on how to implement the following row rendering. I don't need actual code but just the concept. Row 1 is Yellow and Row 2 is White. My bean has Order Date, Req Ship Date, ... and Product as properties.

Order Date	Req Ship Date	InHand Date	Customer	Ship Date	WO#	Ship By	Ready	Shipped	Notes
Product:									
Product:									

B. Keywords: java, swing (java), table (database)

Q. Please rate how well the keywords describe the question in your opinion (Section B)

Very bad ☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 **Very good**

59. **A. local storage list implementation**

Is there a implementation of List as a local storage instead of main memory resident in java. Basically my application have to store a huge amount of data and i don't want to use the memory resident list implementations.

B. Keywords: java, main memory, computer data storage, data

Q. Please rate how well the keywords describe the question in your opinion (Section B)

Very bad ☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 **Very good**

60. **A. local storage list implementation**

Is there a implementation of List as a local storage instead of main memory resident in java. Basically my application have to store a huge amount of data and i don't want to use the memory resident list implementations.

B. Keywords: java, list

Q. Please rate how well the keywords describe the question in your opinion (Section B)

Very bad ☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 **Very good**

A.2 Answers Experiment

Answers experiments showed participants 46 questions. Each question consisted of firstly, an unanswered question, secondly, a question similar to the first unanswered question that had an answer that can answer the first question and finally, the second question's answer that could provide a solution to the first question. Participants were asked to rate how well the answer to provide a solution to the first question. Participants rated the answers on the scale of 1 to 10 where 1 was very bad and 10 was very good. 15 questions were asked from programming language Python³ and Java programming language experiment was broken into two sets. First one⁴ consisting of 15 questions and the second one⁵ consisting of 16 questions. All the 46 questions with instructions are below.

Instruction for question:

1. Read the question in both section A and B
2. Read section C which is the answer to B
3. Rate how well the answer in section C also answers the question in section A in your opinion
- 4 . You are allowed to click on any links if you want.
5. Please don't Google the question as it might show you StackOverflow results and its rating might influence your rating.

1. A. Read text from PNG with standard lib

Is there a way to read text from a PNG-File in Python by using only the standard libraries Python provides?

Similar Question:

B. Python: default/common way to read png images

I haven't found a standard way in Python to read images. Is there really none (because there are so many functions for so many custom stuff that I really wonder that there are no functions to read images)? Or what is it? (It should be available in the MacOSX standard installation and in most recent versions on Linux distributions.)

If there is none, what is the most common lib?

³<https://www.isurvey.soton.ac.uk/start.php?id=13709>

⁴<https://www.isurvey.soton.ac.uk/start.php?id=57376>

⁵<https://www.isurvey.soton.ac.uk/start.php?id=57853>

Many search results hint me to Python Imaging Library. If this is some well known Python-lib for reading images, why isn't it included in Python?

Answer:

C. No, there are no modules in the standard library for reading/writing/processing images directly. But the most common library might be PIL (Python Imaging Library). Many projects are not included in the standard library because they are 1) totally optional and 2) cannot be maintained by the few Python core developers.

Q. Please rate how well the similar question and answer (section B and C) provide a solution to the first question (section A).

Very bad ☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ☐ 6 ☐ 7 ☐ 8 ☐ 9 ☐ 10
Very good

2. A. How to programmatically create dummy email thread?

For comparing some software I need a thread of mails, i.e. some mails with replies and replies to the replies... Content does not matter, but attachment and richt-text would be nice.

I wonder how to create such a dummy mail thread programmatically (preferably using Linux commandline tools or Python).

How would I create those dummy mails?

Similar Question:

B. How to design an email system?

I am working for a company that provides customer support to its clients. I am trying to design a system that would send emails automatically to clients when some event occurs. The system would consist of a backend part and a web interface part. The backend will handle the communication with a web interface (which will be only for internal use to change the email templates) and most important it will check some database tables and based on those results will send emails ... lots of them.

Now, I am wondering how can this be designed so it can be made scalable and provide the necessary performance as it will probably have to handle a few thousands emails per hours (this should be the peek). I am mostly interested about how would this kind of architecture should be thought in order to be easily scaled in the future if needed.

Python will be used on the backend with Postgres and probably whatever comes first between a Python web framework and GWT on the frontend (which seems

the simplest task).

Answer:

C. This is a real good candidate for using some off the shelf software. There are any number of open-source mailing list manager packages around; they already know how to do the mass mailings. It's not completely clear whether these mailings would go to the same set of people each time; if so, get any one of the regular mailing list programs. If not, the easy answer is

```
$ mail address -s subject < file
once per mail.
```

By the way, investigate the policies of whoever is upstream from you on the net. Some ISPs see lots of mails as probable spam, and may surprise you by cutting off or metering your internet access.

Q. Please rate how well the similar question and answer (section B and C) provide a solution to the first question (section A).

Very bad ☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ☐ 6 ☐ 7 ☐ 8 ☐ 9 ☐ 10
Very good

3. A. How to call python code from excel?

I have written a python code using selenium for automating files extraction from website and then I need to format these reports and append them and i am using macros for this.

Is there any way to call the python code from excel by integrating it with VBA or something like that?

Similar Question:

B. can i run c# code from vba macro?

is it possible to include some kind of libraries inside of VBA that will enable me to use c# functions that i wrote

Answer:

C. You need to expose managed code to COM using the [ComVisible] attribute. For more information, see [here](#).

EDIT: For example:

```
[ComVisible(true)]
public class MyClass {
    public int GetNumber(string name) { ... }
}
```

Q. Please rate how well the similar question and answer (section B and C) provide a solution to the first question (section A).

Very bad ☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ☐ 6 ☐ 7 ☐ 8 ☐ 9 ☐ 10
Very good

4. A. In Python IDLE show completion what does -¿ mean?

I managed to get Show Completion to work thanks to this answer.
 But what does `str(object) -> string` mean as a tip after typing the opening bracket
 Example code:

```
linkText = "some text"
elms = browser.find_elements(By.PARTIAL_LINK_TEXT(linkText))
On Run gives: TypeError: 'str' object is not callable
```

Does it mean `linkText` should be a pointer to string? How do I enter a pointer in Python?

Similar Question:

B. Python IDLE subprocess error?

IDLE's subprocess didn't make connection. Either IDLE can't start a subprocess or personal firewall software is blocking the connection.

Don't think this has been asked-how come this comes up occasionally when running very simple programs-I then have to go to Task Manager & stop all Pythonw processes to get it to work again?

It seems to happen randomly on different bits of code-here is the one I'm doing at the moment-

```
f = open('money.txt')
currentmoney = float(f.readline())
print(currentmoney, end='')
howmuch = (float(input('How much did you put in or take out?:'))))
now = currentmoney + howmuch
```

```

print(now)
f.close()
f = open('money.txt', 'w')
f.write(str(now))
f.close()

```

Sometimes it works, sometimes it doesn't!

Answer:

C. You can use `idle -n` to avoid such issues (albeit possibly getting some other limitations).

Q. Please rate how well the similar question and answer (section B and C) provide a solution to the first question (section A).

Very bad ☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ☐ 6 ☐ 7 ☐ 8 ☐ 9 ☐ 10
Very good

5. A. python significant figures and float

I'm do not understand why $1.1 + 2.2$ is not 3.3 if a computer calculates this. I am trying to understand the working of binary floating points.. but I am not even sure of that float the cause is. could you explain this to me?, I have not been able to find a clear explanation.

```

Python 2.7.4 (default, Apr  6 2013, 19:54:46) [MSC v.1500 32 bit (Intel)] on
Type "copyright", "credits" or "license()" for more information.
>>> ===== RESTART =====
>>>
>>> 1.1+2.2
3.3000000000000003
>>>

```

Similar Question:

B. Python float - str - float weirdness

```

>>> float(str(0.65000000000000002))

0.65000000000000002

>>> float(str(0.47000000000000003))

```

0.46999999999999997 ???

What is going on here? How do I convert 0.47000000000000003 to string and the resultant value back to float?

I am using Python 2.5.4 on Windows.

Answer:

C. `str(0.47000000000000003)` give '0.47' and `float('0.47')` can be 0.46999999999999997. This is due to the way floating point number are represented (see this wikipedia article) Note: `float(repr(0.47000000000000003))` or `eval(repr(0.47000000000000003))` will give you the expected result, but you should use Decimal if you need precision.

Q. Please rate how well the similar question and answer (section B and C) provide a solution to the first question (section A).

Very bad ☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ☐ 6 ☐ 7 ☐ 8 ☐ 9 ☐ 10
Very good

6. A. How to verify ssl certificate?

im working on dns masking. Im asking user to provide URL for masking and SSL certificate. I need to validate whether the certificate user provided is of valid format or not. The problem is if user doesnt provide valid certificate apache doesn't restart. Im using python.

I've tried some solutions provided here but couldn't validate ssl certificate.

Any help will be highly appreciated.

Thanks!

Similar Question:

B. Validate SSL certificates with Python I need to write a script that connects to a bunch of sites on our corporate intranet over HTTPS and verifies that their SSL certificates are valid; that they are not expired, that they are issued for the correct address, etc. We use our own internal corporate Certificate Authority for these sites, so we have the public key of the CA to verify the certificates against. Python by default just accepts and uses SSL certificates when using HTTPS, so even if a certificate is invalid, Python libraries such as urllib2 and Twisted will just happily use the certificate.

Is there a good library somewhere that will let me connect to a site over HTTPS and verify its certificate in this way?

How do I verify a certificate in Python?

Answer:

C. I have added a distribution to the Python Package Index which makes the `match_hostname()` function from the Python 3.2 `ssl` package available on previous versions of Python. http://pypi.python.org/pypi/backports.ssl_match_hostname/

You can install it with:

```
pip install backports.ssl_match_hostname
```

Or you can make it a dependency listed in your project's `setup.py`.

Either way, it can be used like this:

```
from backports.ssl_match_hostname import match_hostname, CertificateError
...
sslsock = ssl.wrap_socket(sock, ssl_version=ssl.PROTOCOL_SSLv3,
                           cert_reqs=ssl.CERT_REQUIRED, ca_certs=...)
try:
    match_hostname(sslsock.getpeercert(), hostname)
except CertificateError, ce:
    ...
```

Q. Please rate how well the similar question and answer (section B and C) provide a solution to the first question (section A).

Very bad ☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ☐ 6 ☐ 7 ☐ 8 ☐ 9 ☐ 10
Very good

7. A. HTTP Request coming up blank

With my first foray into anything-related-to-web programming, I am using the Requests external library (Requests), I am testing it on <https://google.com.au> to see if I'm doing it right.

```
import requests

proxy_dict = {
    "https" : "10.10.20.99:8080"
}

r = requests.get('https://google.com.au', proxies = proxy_dict)
print r.content
print r.status_code
```

with output

```
<HTML>>
```

```
200
```

i.e. a completed HTTP request yet with no returned HTML information. I have read the relevant docs for "Requests" but I can't get the same results as that example (which is the same as this one, except using github.com instead of google.com.au). I am a complete and utter noob at all things HTTP/HTML right now, so does anyone have some idea where I'm going wrong?

Thanks!

EDIT: I forgot to mention, this will be behind some sort of company firewall as I am doing this through work. That proxy is my work's proxy.!

Similar Question:

B. Python Proxy Error With Requests Library I am trying to access the web via a proxy server in Python. I am using the requests library and I am having an issue with authenticating my proxy as the proxy I am using requires a password.

```
proxyDict = {
    'http' : 'username:mypassword@77.75.105.165',
    'https' : 'username:mypassword@77.75.105.165'
}

r = requests.get("http://www.google.com", proxies=proxyDict)
I am getting the following error:
```

```
Traceback (most recent call last):
```

```
File "", line 1, in <module>
```

```
r = requests.get("http://www.google.com", proxies=proxyDict)
```

```
File "C:\Python27\lib\site-packages\requestsapi.py", line 78, in get
```

```
:param url: URL for the new :class:'Request' object.
```

```
File "C:\Python27\lib\site-packages\requestsapi.py", line 65, in request
```

```
"""Sends a POST request. Returns :class:'Response' object.
```

```
File "C:\Python27\lib\site-packages\requestssessions.py", line 187, in request
```

```
def head(self, url, **kwargs):
```

```
File "C:\Python27\lib\site-packages\requestsmodels.py", line 407, in send
```

```
"""
```

```
File "C:\Python27\lib\site-packages\requestspackagesurllib3poolmanager.py", line 127, i
```

```
File "C:\Python27\lib\site-packages\requestspackagesurllib3connectionpool.py", line 521
```

```
File "C:\Python27\lib\site-packages\requestspackagesurllib3connectionpool.py", line 497
```

```
ValueError: invalid literal for int() with base 10: 'h6f2v6jh5dsxa@77.75.105
```

How do I solve this?

Thanks in advance for your help.

Answer:

C. You should remove the embedded username and password from proxyDict, and use the auth parameter instead.

```
import requests
from requests.auth import HTTPProxyAuth

proxyDict = {
    'http' : '77.75.105.165',
    'https' : '77.75.105.165'
}

auth = HTTPProxyAuth('username', 'mypassword')

r = requests.get("http://www.google.com", proxies=proxyDict, auth=auth)
```

Q. Please rate how well the similar question and answer (section B and C) provide a solution to the first question (section A).

Very bad ☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ☐ 6 ☐ 7 ☐ 8 ☐ 9 ☐ 10
Very good

8. A. Python Java Integration

I'm developing a program completely written in Python but I need to integrate a Java code inside the file console.py. I want to integrate Sphinx4's program to give GNS3 the capability of voice recognition.

Is it possible? What do I need to do this!

Similar Question:

B. Java Python Integration I have a Java app that needs to integrate with a 3rd party library. The library is written in Python, and I don't have any say over that. I'm trying to figure out the best way to integrate with it. I'm trying out JEPP (Java Embedded Python) - has anyone used that before? My other thought is to use JNI to communicate with the C bindings for Python.

Any thoughts on the best way to do this would be appreciated. Thanks.

Answer:

C. Why not use Jython? The only downside I can immediately think of is if your library uses CPython native extensions. EDIT: If you can use Jython now but think you may have problems with a later version of the library, I suggest you try to isolate the library from your app (e.g. some sort of adapter interface). Go with the simplest thing that works for the moment, then consider JNI/CPython/etc if and when you ever need to. There's little to be gained by going the (painful) JNI route unless you really have to.

Q. Please rate how well the similar question and answer (section B and C) provide a solution to the first question (section A).

Very bad ☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ☐ 6 ☐ 7 ☐ 8 ☐ 9 ☐ 10
Very good

9. **A. What is the most painless way to remove one version of python and not the other on OSX 10.7.4?**

I accidentally installed 64 bit python a few weeks ago, and now all of my programs jump for it by default, instead of the more functional 32 bit python. Is there a painless way to uninstall 64 bit python but not 32?

Thanks!

Similar Question:

B. How to uninstall Python 2.7 on a Mac OS X 10.6.4? I want to completely remove Python 2.7 from my Mac OS X 10.6.4. I managed to remove the entry from the PATH variable by reverting my .bash_profile. But I also want to remove all directories, files, symlinks, and entries that got installed by the Python 2.7 install package. I've got the install package from <http://www.python.org/>. What directories/files/configuration file entries do I need to remove? Is there a list somewhere?

Answer:

C. The complete list is documented here. But, basically, all you need to do is to: remove the Python 2.7 framework

```
sudo rm -rf /Library/Frameworks/Python.framework/Versions/2.7
```

remove the Python 2.7 applications directory

```
sudo rm -rf "/Applications/Python 2.7"
```

remove the symbolic links in /usr/local/bin that point to this python version see `ls -l /usr/local/bin | grep '.*Library/Frameworks/Python.framework/Versions/2.7'`

Q. Please rate how well the similar question and answer (section B and C) provide a solution to the first question (section A).

Very bad ☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ☐ 6 ☐ 7 ☐ 8 ☐ 9 ☐ 10
Very good

10. A. Free memory in python run time

I have a python server running on tornado , there is request that uses lots of memory (using a gdata library) . I deleted those object after using them and even did `gc.collect()` , But when i see the system memory using `free -m` , I could see memory increase in memory when i do those operations using gdata . But the memory does not get freed up , after i delete the object . The memory gets freed up only after the main python program is killed . I want to know if there is any way to free up the memory .

Similar Question:

B. Python is not freeing memory I've been working with XML resources, and it seems that Python is issuing a weird behavior. I've tested both `lxml` library and `xml.etree.ElementTree`, both holding memory after it should be collected by `gc`. I typed `gc.collect()` as a test, but nothing else happen: memory still beign hold by process.

Imports:

```
import time
from lxml import etree
import gc
This is the code:
```

```
def process_alternative():
    """
    This alternative process will use lxml
    """
    filename = u"/tmp/randomness.xml"
    fd = open(filename, 'r')
    tree = etree.parse(fd)
    root = tree.getroot()

    accum = {}
```

```

    for _item in root.iter("*"):
        for _field in _item.iter("*"):
            if _field.tag in accum.keys():
                accum[_field.tag] += 1
            else:
                accum[_field.tag] = 1

    for key in accum.keys():
        print "%s -> %i" % (key, accum[key])

    fd.close()
    gc.collect()
And this is my main

if __name__ == "__main__":
    while True:
        print "Wake up!"
        process_alternative()
        print "Sleeping..."
        time.sleep(30)

```

As you see, this main calls "process_alternative", and then sleep. XML file provided loads memory with nearly 800Mb; so, before time.sleep, memory should be freed by process, returning to basic VM memory needed (around 32Mb?). Instead, process continue holding around 800Mb.

Any tip about why memory has not been freed after every iteration?

Thank you so much,

EDIT:

Using ubuntu 13.04, Python 2.7.4

EDIT2:

This function deallocates memory in every iteration

```

def check_memory():
    ac1 = [a1**5 for a1 in xrange(10000000)]
    time.sleep(5)
    ac2 = [a1**5 for a1 in xrange(10000000)]
    time.sleep(5)
    ac3 = [a1**5 for a1 in xrange(10000000)]

```

Answer:

C. I do not know why, but process still hold memory, even when I set an explicit call to `gc.collect()`. After some playing, and thanks to Martijn Pieters, a solution appeared. Calling

```
len(gc.get_objects())
```

frees all accessed memory, and keeps process on right resources when it's not busy. Strange, but true.

Thank you for all responses

Cheers,

Isaac

Q. Please rate how well the similar question and answer (section B and C) provide a solution to the first question (section A).

Very bad ☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ☐ 6 ☐ 7 ☐ 8 ☐ 9 ☐ 10
Very good

11. **A. URL links in Eclipse python code**

I like to document my python code with links to web pages that explain the algorithm in detail. But I can't make the links 'live', so that just clicking on them in Eclipse brings up the page in the platform browser.

I've just been putting the URL's inside the 3-double-quote pair, then having to manually copy paste them to the browser.

Is there a better way?

Similar Question:

B. Python DocStrings **Pydev** I've gotten Pydev up and running, and almost all is working well. However I'm having some trouble with docstrings.

Let's say for instance I have a function such as the following:

```
def _get_logging_statement(self):
    """Returns an easy to read string which separates items in the log file
    result = "nn#======"
    result += "n#    %-80s#"%(self)
    result += "nn#======"
    return result
```

Assume I've overridden `repr` to format that string properly as well.

When I hover over this in Eclipse it's showing me the full docstring as intended, however below the docstring is the implementation. Is there a way to show only the docstring?

Answer:

C. Doesn't look like it currently. Googled around for this issue and the top result pointed me to this Pydev-users post:

On Mon, May 3, 2010 at 5:45 AM, Janosch Peters wrote:

Hi,

when I hover over a function or class, I get a tooltip showing the whole definition of the function/class not only the docstring (as I would expect).

Is this expected behaviour? I think it would be more useful, if only the content of the docstring is shown.

It's currently expected. Please enter a feature request to make showing just the docstring an option.

Cheers,

Fabio

Looked around the Pydev bug/feature tracker and didn't find this specific issue entered. You might want to enter it in the Pydev feature request tracker and see if you can get help there.

Q. Please rate how well the similar question and answer (section B and C) provide a solution to the first question (section A).

Very bad ☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ☐ 6 ☐ 7 ☐ 8 ☐ 9 ☐ 10
Very good

12. A. need a time widget or clock widget

is there anything like a timewidget or clock widget in Django which will help me to input data into a form for appointment date (like we have calendar or date widget).

i have my date widget working with the below code in my forms.py:

```
import datetime
from django.forms.extras.widgets import SelectDateWidget
mydate = forms.DateField(widget=SelectDateWidget)
```


is there anything like this for time widget

Similar Question:

B. Using Django time/date widgets in custom form How can I use the nifty JavaScript date and time widgets that the default admin uses with my custom view?

I have looked through the Django forms documentation, and it briefly mentions `django.contrib.admin.widgets`, but I don't know how to use it?

Here is my template that I want it applied on.

```
<form action="." method="POST">
  <table>
    {% for f in form %}
      <tr> <td> {{ f.name }}<td> {{ f }}<td> tr>
    {% endfor %}
  </table>
  <input type="submit" name="submit" value="Add Product">
</form>
```

Also, I think it should be noted that I haven't really written a view up myself for this form, I am using a generic view. Here is the entry from the `url.py`:

```
(r'^admin/products/add/$', create_object, {'model': Product, 'post_save_redirect': ...})
```

And I am relevantly new to the whole Django/MVC/MTV thing, so please go easy...

Answer:

C. What about just assigning a class to your widget and then binding that class to the JQuery datepicker?

Django `forms.py`:

```
class MyForm(forms.ModelForm):

    class Meta:
        model = MyModel

    def __init__(self, *args, **kwargs):
        super(MyForm, self).__init__(*args, **kwargs)
```

```
self.fields['my_date_field'].widget.attrs['class'] = 'datepicker'
```

And some JavaScript for the template:

```
$(".datepicker").datepicker(); \\\
```

Q. Please rate how well the similar question and answer (section B and C) provide a solution to the first question (section A).

Very bad ☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ☐ 6 ☐ 7 ☐ 8 ☐ 9 ☐ 10

Very good

13. A. Python child process limits

I'd like to create a process in Python (probably with subprocess and Popen), which should have limited CPU time, limited child processes and memory bandwidth. I can't find a way to do this. resource.setrlimit does not seem to work.

My code is :

```
import os
import sys
import resource
import subprocess
import signal

def setlimits():
    os.seteuid(65534) # Has to run as root user in order to be able to setuid
    resource.setrlimit(resource.RLIMIT_CPU, (1, 1))
    resource.setrlimit(resource.RLIMIT_FSIZE, (500, 500))
    resource.setrlimit(resource.RLIMIT_NPROC, (80, 80))

p = subprocess.Popen( ["./exec.out"] , preexec_fn=setlimits )
```

Similar Question:

B. Python: ulimit and nice for subprocess.call / subprocess.Popen? I need to limit the amount of time and cpu taken by external command line apps I spawn from a python process using subprocess.call , mainly because sometimes the spawned process gets stuck and pins the cpu at 99

nice and ulimit seem like reasonable ways to do this, but I'm not sure how they'd interact with subprocess.

The limits look something like: Kill the process if it's taking more than 60 seconds Limit it to 20I want to apply the resource limiting to the subprocess, not to the

python process that's spawning the subprocesses. Is there a way to apply nice and ulimit to the subprocess.call spawned process? Are there better python-native alternatives?

This is on a linux (ubuntu) system.

Answer:

C. You can set limits for subprocesses with the ulimit and nice shell commands like this: `import subprocess subprocess.Popen('ulimit -t 60; nice -n 15 cpuhog', shell=True)` This runs cpuhog with a limit of 60 seconds of CPU time and a niceness adjustment of 15. Note that there is no simple way to set a 20% CPU throttle as such. The process will use 100% CPU unless another (less nice) process also needs the CPU.

Q. Please rate how well the similar question and answer (section B and C) provide a solution to the first question (section A).

Very bad ☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ☐ 6 ☐ 7 ☐ 8 ☐ 9 ☐ 10
Very good

14. A.Using C++ library functions from Python

I got a library created for Win x86 with MS Visual Studio 2010. And I can not change the content of a library to use Boost.Python.

I'm using Python 3.3 with PyQt4 to create interface, but not restricted to these versions. I need to call functions and get objects from said C++ library. What is the easiest way to wrap C++ library to be called from python?

I guess, that such question was already asked, but I can not seem to find it.

Here's an example of header file:

```
namespace SDK
{ class IMethod
{
public:
    virtual IModel* CreateModel(const IBuffer* pBuffer, const char* text)
};
extern __declspec(dllexport) SDK::IMethod* CreateMethod(MethodID integer)
}
```

Similar Question:

B. Python — accessing dll using ctypes I'm trying to access some functions in a dll (nss3.dll) that ships with Firefox web browser. To handle this task I have

used ctypes in Python. The problem is that it fails at the initial point which is when loading the dll in to the memory.

This is the code snippet that I have to do so.

```
>>> from ctypes import *
>>> windll.LoadLibrary("E:nss3.dll")
The exception I'm getting is

Traceback (most recent call last):
  File "", line 1, in <module>
    windll.LoadLibrary("E:nss3.dll")
  File "C:Python26libctypes__init__.py", line 431, in LoadLibrary
    return self._dlltype(name)
  File "C:Python26libctypes__init__.py", line 353, in __init__
    self._handle = _dlopen(self._name, mode)
WindowsError: [Error 126] The specified module could not be found
I also tried loading it from the Firefox
installation path assuming that there maybe dependencies.

>>> windll.LoadLibrary("F:SoftwaresMozilla Firefoxnss3.dll")
```

But I'm getting the same exception as mentioned above.

Thanks.

Answer:

C. nss3 is linked to several DLLs in the Firefox directory: nssutil3.dll, plc4.dll, plds4.dll, nspr4.dll, and MOZCRT19.dll. You need to add this directory to the system PATH:

```
import os
import ctypes

firefox = r'F:SoftwaresMozilla Firefox'
os.environ['PATH'] = ';' + os.path.join(firefox, os.environ['PATH'])
Then more than likely you'll want to use cdll for the cdecl calling convention:

>>> nss3 = ctypes.CDLL(os.path.join(firefox, 'nss3.dll'))

>>> nss3.NSS_GetVersion.restype = c_char_p
>>> nss3.NSS_GetVersion()
'3.13.5.0 Basic ECC'
```

Q. Please rate how well the similar question and answer (section B and C) provide a solution to the first question (section A).

Very bad ☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ☐ 6 ☐ 7 ☐ 8 ☐ 9 ☐ 10
Very good

15. **A. python pprint: how to output utf8 characters?**

I have a dictionary with 'utf8 char' : number . pprint() would treat the utf8 as a byte array and output hex values. Is it possible to tell it to print it as string so that the console can render UTF8 text?

Similar Question:

B. Unable to encode/decode pprint output This question is based on a side-effect of that one.

My .py files are all have `-*- coding: utf-8 -*-` encoding definer on the first line, like my api.py

As I mention on the related question, I use `HttpResponse` to return the api documentation. Since I defined encoding by:

`HttpResponse(cy_content, content_type='text/plain; charset=utf-8')` Everything is ok, and when I call my API service, there are no encoding problems except the string formed from a dictionary by pprint

Since I am using Turkish characters in some values in my dict, pprint converts them to unichr equivalents, like:

```
API_STATUS = {
    1: 'müşteri',
    2: 'some other status message'
}
```

```
my_str = 'Here is the documentation part that contains Turkish chars like işüğçö'
my_str += pprint.pformat(API_STATUS, indent=4, width=1)
return HttpResponse(my_str, content_type='text/plain; charset=utf-8')
And my plain text output is like:
```

```
Here is the documentation part that contains Turkish chars like işüğçö

{
    1: 'mxc3xbxc5x9fteri',
```

```

    2: 'some other status message'
}

```

I try to decode or encode pprint output to different encodings, with no success...
What is the best practice to overcome this problem

Answer:

C. pprint appears to use repr by default, you can work around this by overriding
PrettyPrinter.format:

```

# coding=utf8
import pprint

class MyPrettyPrinter(pprint.PrettyPrinter):
    def format(self, object, context, maxlevels, level):
        if isinstance(object, unicode):
            return (object.encode('utf8'), True, False)
        return pprint.PrettyPrinter.format(self, object, context, maxlevels, level)

d = {'foo': u'ışüğçö'}

pprint.pprint(d)          # {'foo': u'iu015fxfcu011fxe7xf6'}
MyPrettyPrinter().pprint(d) # {'foo': ışüğçö}

```

Q. Please rate how well the similar question and answer (section B and C) provide
a solution to the first question (section A).

Very bad ☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ☐ 6 ☐ 7 ☐ 8 ☐ 9 ☐ 10
Very good

16. **A. Java, console.readPassword adds extra line. How to delete it?**

When i use console.readPassword() to read user passwords through console, there
is always one line added to the console.

How to disable this behavior or how to delete that extra line (and move the cursor
after the last character in the line before)? What escape character to use?

Thanks

Similar Question:

B. Java console pretty printing: returning to the start of a line In Java, how can I return to the start of a line and overwrite what has already been output on the console?

`System.out.print(mystuff+'r');` does not appear to work.

thanks in advance

Answer:

C. I suspect that your cursor IS moving to the front of the line. The text you already have isn't disappearing because you haven't overwritten it with anything. You could output spaces to blank the line and then add another r. I just tested the following on Windows XP and AIX and it works as expected:

```
public class Foo {
    public static void main(String[] args) throws Exception {
        System.out.print("old line");
        Thread.sleep(3000);
        System.out.print("rnew");
    }
}
```

I get "old line" printed, a 3 second delay, and then "old line" changes to "new line"

I intentionally made the first line longer than the second to demonstrate that if you want to erase the whole line you'd have to overwrite the end with spaces.

Also note that the "b" escape sequence will back up 1 space, instead of to the beginning of the line. So if you only wanted to erase the last 2 characters, you could write:

`System.out.println("foobbun")` and get "fun".

Q. Please rate how well the similar question and answer (section B and C) provide a solution to the first question (section A).

Very bad ☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ☐ 6 ☐ 7 ☐ 8 ☐ 9 ☐ 10
Very good

17. A. is there an equivalent of Python's timeit module in Java stdlib

I was amazed at Python's timeit module and wondered if there's an equivalent in Java's standard library. If not, is there a 3rd party module?

Similar Question:

B. How to compute accurately the time it takes a Java program to write or read a file? How to compute accurately the time it takes a Java program to write or read a number of bytes from/to a file ?

It is really important that the time is being measured accurately. (The time should be computed by the program itself).

Answer:

C. The standard idiom is: `long startTime = System.nanoTime(); doSomething();`
`long elapsedTime = System.nanoTime() - startTime;`

Q. Please rate how well the similar question and answer (section B and C) provide a solution to the first question (section A).

Very bad ☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ☐ 6 ☐ 7 ☐ 8 ☐ 9 ☐ 10
Very good

18. **A. java.lang.NoSuchMethodError when trying to read .xslm file**

Exception in thread "main" java.lang.NoSuchMethodError: org.apache.xmlbeans.XmlOptions.setSaveAgg at org.apache.poi.POIXMLDocumentPart.<clinit>(POIXMLDocumentPart.java:56) at rulebooksToExcel.GenerateExcel.generateExcel(GenerateExcel.java:34) at rulebooksToExcel.ParseNortDocFiles.main(ParseNortDocFiles.java:165) I am getting the error at :

`workbook = new XSSFWorkbook(in);` I read other similar questions but they all suggest XMLBeans Version 2.0+. But I am using 2.6, and I can't find any other explanation for what might be causing this.

Similar Question:

B. How to solve a NoSuchMethodError when using POI for doc files

When I was trying to implement any code of the following

```
File someFile = new File("D:arz.doc");
InputStream inputStrm = new FileInputStream(someFile);
HWPFDocument wordDoc = new HWPFDocument(inputStrm);
System.out.println(wordDoc.getText());
or:
```

```
POIFSFileSystem fs = new POIFSFileSystem(new FileInputStream("D:arz.doc"));
WordExtractor extractor = new WordExtractor(fs);
String wordText = extractor.getText();
```


, the error message always comes out as following:

```
Exception in thread \main" java.lang.NoSuchMethodError:
org.apache.poi.poifs.filesystem.POIFSFileSystem.getRoot()Lorg/apache/poi/poifs/
at org.apache.poi.hwpf.HWPFDocument.(HWPFDocument.java:186)
at DB_connect.dissertation_aralaz.ParseWodDocFile.main(ParseWodDocFile.java:2
Java Result: 1
BUILD SUCCESSFUL (total time: 3 seconds)
Could you please help me in that problem?\\
```

Answer:

C. You almost certainly have two copies of POI on your classpath. One is the new, latest version which contains the feature you want to use. The other is an older version that doesn't, and it seems your system is preferring the older one... This is a common enough problem that the POI FAQ Covers this very case. Ideally, just look at your classpath, and try to identify the extra older POI jar. However, if that doesn't work, try this snippet of code from the POI FAQ:

```
ClassLoader classloader =
    org.apache.poi.poifs.filesystem.POIFSFileSystem.class.getClassLoader();
URL res = classloader.getResource(
    "org/apache/poi/poifs/filesystem/POIFSFileSystem.class");
String path = res.getPath();
System.out.println("Core POI came from " + path);
```

That will print out the filename of the POI jar you're using, so you can work out where the older copy is coming from and remove it!

Q. Please rate how well the similar question and answer (section B and C) provide a solution to the first question (section A).

Very bad ☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ☐ 6 ☐ 7 ☐ 8 ☐ 9 ☐ 10
Very good

19. A. How do we call a method in a jar file from a java script method

How do we call a method in a jar file from a java script method. I am using a third party to authenticate a HTML5 application, the jquery method redirects to the third party url and they redirect back to the app after validation at their end.

Now in Java we can use the methods in their jar file to get back the User ID, but I am not sure if we can do it using js.

Code in java, the jar is added in the classpath -

```
private static String UID(HttpServletRequest req) {
    String unEncCookie = null;
    String cookie = getSecCookie(req);
    if (cookie == null)
        return null;
    else {
        unEncCookie = JAR.JAR(cookie, "param1", "param2");
        if (unEncCookie == null || "".equals(unEncCookie))
            return null;
        else{
            return unEncCookie.split("|")[5]; // 6th value is UID
        }
    }
}
```

Similar Question:

B. calling java methods in javascript code i created a java class content method return a String, my question is how to call this function in my javascript code to use the returned value from the java method. I want to call client-side Java code embedded in browser.

here is an exemple of what im talking about:

in my webpage i have a javascript code, here is some of it:

```
function createChartControl(htmlDiv1)
{
    // Initialize Gantt data structures
    //project 1
    var parentTask1 = new GanttTaskInfo(1, "Old code review", new Date(2010, 5, 11))
    .....
```

i want to create a java class content methods to provide data to this javascript function "GanttTaskInfo". for exemple function to get name, get id and date. well i think this time im clear :D i searched a way to call java methods in javascript, and i found applets as you said, but i think its not usefull to me. thanks again

Answer:

C. Java is a server side language, whereas javascript is a client side language. Both cannot communicate. If you have setup some server side script using Java you could use AJAX on the client in order to send an asynchronous request to it and thus invoke any possible Java functions. For example if you use jQuery as js framework you may take a look at the \$.ajax() method. Or if you wanted to do it using plain javascript, here's a tutorial.

Q. Please rate how well the similar question and answer (section B and C) provide a solution to the first question (section A).

Very bad ☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ☐ 6 ☐ 7 ☐ 8 ☐ 9 ☐ 10
Very good

20. **A. Is there any Java API for PCRE regex matching**

we have a requirement to match the PCRE expression against a file name. Tried to find the API regarding it, but unable to find any. Need some suggestion how to achieve it. If any of you have already done, a sample would be more grateful

Similar Question:

B. What support is there for PCRE (Perl Compatible Regular Expressions) in common languages? I am interested in the power of PCRE (Perl Compatible Regular Expressions) and wonder whether they are likely to become a de facto approach in all major languages (I am interested in Java). I am prepared to use a library if necessary.

I also could not find a good page in SO describing the pros and cons of PCRE so if this does not exist it could be useful to include this in answers

EDIT I am interested in power beyond Java 1.6 regex, particularly named capture groups

Answer:

C. It seems that more mainstream languages actually use their own implementation of "Perl-like" regexes than actually use libpcre. Languages that fall into this class include (at the very least) Java, JavaScript, and Python. Java's java.util.regex library uses a syntax that's very heavily based on Perl (approx. version 5.8) regexes, including the rules for escaping, the p and P Unicode classes, non-greedy and "possessive" quantifiers, backreferences, Q..E quoting, and several of the (?...) constructs including non-capturing groups, zero-width lookahead/behind, and non-backtracking groups. In fact Java regexes seem to have more in common with Perl regexes than libpcre does. :)

The JavaScript language also uses regexes that are derived from Perl; Unicode classes, lookbehind, possessive quantifiers, and non-backtracking groups are absent, but the rest of what I mentioned for Java is present as well in JS.

Python's regex syntax is also based on Perl 5's, with non-greedy quantifiers, most of the (?...) constructs including non-capturing groups, look-ahead/behind and conditional patterns, as well as named capture groups (but with a different syntax than either Perl or PCRE). Non-backtracking groups and 'possessive' quantifiers are (as far as I can see) absent, as are p and P Unicode character classes, although the standard d, s, and w classes are Unicode-aware if requested.

Q. Please rate how well the similar question and answer (section B and C) provide a solution to the first question (section A).

Very bad ☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ☐ 6 ☐ 7 ☐ 8 ☐ 9 ☐ 10
Very good

21. **A. Play Framework 2.1.3 Debugger failed to attach: handshake failed**

I get this message in the console when I run a Play.

```
C:\Users\Daniel\desenvolvimento\Sistemas>play debug
Listening for transport dt_socket at address: 9999
Debugger failed to attach: handshake failed - received >POST /setwindo<-
expected >JDWP-Handshake<
Debugger failed to attach: handshake failed - received >POST /setwindo<-
expected >JDWP-Handshake<
```

I don't have any problems running the app but this "Debugger failed" message keeps coming out and it just bugs me.

This happens even if I create a clean project.

Play 2.1.3 Windows7 64bit

How could I get rid of this message?

Similar Question:

B. Play Framework 1.2.3 Debugger failed to attach: handshake failed

I get this message in the console when I run a Play app.

```
06:08:08,069 INFO ~ Starting D:\projects\play1.2.3\test
06:08:08,623 WARN ~ You're running Play! in DEV mode
06:08:08,737 INFO ~ Listening for HTTP on port 9000 (Waiting a first request to st
```

```
06:08:31,229 INFO ~ Application 'test' is now started !
Debugger failed to attach: handshake failed - connection prematurely closed
Debugger failed to attach: handshake failed - connection prematurely closed
...
```

I don't have any problems running the app but this "Debugger failed" message keeps coming out and it just bugs me.

This happens even if I create a clean project. ports 8000 and 9000 are open.

System is Java 1.6.0_24 Play 1.2.3 Windows7 64bit

How could I get rid of this message?

Answer:

C. sorry this had nothing to do with play. There seems to be some other connection coming from my router accessing port 8000 and that was the problem.

If I change the jpda.port to something other than 8000 the message doesn't show.

Q. Please rate how well the similar question and answer (section B and C) provide a solution to the first question (section A).

Very bad ☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ☐ 6 ☐ 7 ☐ 8 ☐ 9 ☐ 10
Very good

22. A. Retrieve a list of all packages in a project with Google reflections library?

I'm talking about the reflections lib. Is there any possibility to get a list of all packages which are included in the project where I let the code compile?

I've tried it with the following code bracket but I don't want to insert a project name.

```
Reflections reflections = new Reflections(/* Project name here... */);
Set<Class extends Object>> allClasses =
    reflections.getSubTypesOf(Object.class);
```

Similar Question:

B. Can you find all classes in a package using reflection?

A beginner question about reflection, I suppose:

Is it possible to find all classes or interfaces in a given package? (Quickly looking at e.g. Package, it would seem like no.)

Answer:

C. Due to the dynamic nature of class loaders, this is not possible. Class loaders are not required to tell the VM which classes it can provide, instead they are just handed requests for classes, and have to return a class or throw an exception. However, if you write your own class loaders, or examine the classpaths and it's jars, it's possible to find this information. This will be via filesystem operations though, and not reflection. There might even be libraries that can help you do this.

If there are classes that get generated, or delivered remotely, you will not be able to discover those classes.

The normal method is instead to somewhere register the classes you need access to in a file, or reference them in a different class. Or just use convention when it comes to naming.

Addendum: The Reflections Library will allow you to look up classes in the current classpath. It can be used to get all classes in a package:

```
Reflections reflections = new Reflections("my.project.prefix");

Set<Class extends Object>> allClasses =
    reflections.getSubTypesOf(Object.class);
```

Q. Please rate how well the similar question and answer (section B and C) provide a solution to the first question (section A).

Very bad ☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ☐ 6 ☐ 7 ☐ 8 ☐ 9 ☐ 10
Very good

23. A. Choose Gradle JDK in IntelliJ

My system JAVA_HOME and PATH are pointing to JDK 7, but I want to use JDK 8 in project opened in IntelliJ. I changes settings in Project Structure and it works great in IDE, but unfortunately Gradle build run from IDE still uses JDK 7. How can I specify Gradle JDK in IntelliJ 13.0?

Similar Question:

B. How do I tell IntelliJ to start gradle with Java 1.6 JDK?

A simple question I couldn't figure out even after hours of trying and searching:

I have both Java 6 and 7 installed.

How do I tell IntelliJ to start Gradle builds with a JDK version 1.6?

No matter what I do, IntelliJ keeps starting my gradle with

"c:\program files\java\jdk1.7.0_10\jre\bin\java" -Dgradle.home=C:_CODING_gradle-1.4
[...] I tried:

all project modules are set to use 1.6

JAVA_HOME is set to C:\Program Files\Java\jdk1.6.0_38

setting IDEA_JDK did nothing

gradle -v recognizes 1.6 as its JDK on command line

restarting IntelliJ and/or computer didn't change the behavior

Is there a setting somewhere to avoid Java 1.7 from being invoked?

Answer:

C. 'jre choose' algorithm is defined here, namely, it tries the following:

Java location referenced at 'gradle.java.home' system property (if any);

'JAVA_HOME' env variable;

Project jdk;

Any jdk configured at the ide;

I suspect you have 'gradle.java.home' which points to java 1.7 left at \$IDEA_HOME/bin/idea.properties

Q. Please rate how well the similar question and answer (section B and C) provide a solution to the first question (section A).

Very bad ☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ☐ 6 ☐ 7 ☐ 8 ☐ 9 ☐ 10

Very good

24. A. Serve Images from a separate location in JSF1.1

We are having a lot of images which all exist in our webroot. Is there a setting in JSF 1.1 which allows us to set location of image path.

Similar Question:

B. Load images from outside of webapps / webcontext / deploy folder using h:graphic Image or img tag

I need to display images which reside outside of deploy folder in web application using JSF h:graphicImage tag or HTML img tag. How can I achieve that?

Answer:

C. Several ways: If you have full control over the images folder, then just drop the folder with all images, e.g. /images directly in servletcontainer's deploy folder, such as the /webapps folder in case of Tomcat and /domains/domain1/applications in case of Glassfish. No further configuration is necessary.

Add a new to the servletcontainer with a docBase which points to the absolute disk file system location of the folder with those images. How to do that depends on the container used. In case of for example Tomcat, that'll be the following new entry in /conf/server.xml:

```
docBase="/path/to/images" path="/images" />
```

In case of for example Glassfish, that'll be the following entry in glassfish-web.xml:

```
name="alternatedocroot\_1" value="from=/images/* dir=/path/to" />
```

Create a Servlet which streams the image from disk to response:

```
@WebServlet("/images/*")
```

```
public class ImageServlet extends HttpServlet {
```

```
    protected void doGet(HttpServletRequest request, HttpServletResponse response) {
        String filename = request.getPathInfo().substring(1);
        File file = new File("/path/to/images", filename);
        response.setHeader("Content-Type", getServletContext().getMimeType(filename));
        response.setHeader("Content-Length", String.valueOf(file.length()));
        response.setHeader("Content-Disposition", "inline; filename=\"" + filename + "\"");
        Files.copy(file.toPath(), response.getOutputStream());
    }
}
```

```
}
```

For the first way and the Tomcat approach in second way, the images will be available by `http://example.com/images/filename.ext` and thus referencable in plain HTML as follows

`src="/images/filename.ext" />` For the GlassFish approach in second way and the third way, the images will be available by `http://example.com/context/images/filename.ext` and thus referencable in plain HTML as follows

`src="#request.contextPath/images/filename.ext" />` or in JSF as follows (context path is automatically prepended)

`value="/images/filename.ext" />` See also: recommended way to save files uploaded to a Tomcat servlet Simplest way to serve static data from outside the

application server in a Java web application

Q. Please rate how well the similar question and answer (section B and C) provide a solution to the first question (section A).

Very bad ☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ☐ 6 ☐ 7 ☐ 8 ☐ 9 ☐ 10
Very good

25. **A. Avoid direct links to login page**

I'm using spring security including remember-me feature:

Now when i check remember-me checkbox, i can close browser and go to some app pages skipping the login page.

But i still can type myApp/login.jsp and go to the login page.

Is there any ways to avoid it? - I want to avoid any direct links to login page. User should be able to see login screen only if he hasn't logged in still, hasn't pressed "remember-me" button (and closed browser) or pressed logout button of my app.

Similar Question:

B. Spring Security - Redirect if already logged in

I'm new to Spring:

I do not want authenticated user from accessing the login page. What is the proper way to handle redirects for the '/login' if the user is already authenticated? Say, I want to redirect to '/index' if already logged in.

I have tried 'isAnonymous()' on login, but it redirects to access denied page.

```
auto-config="true" use-expressions="true" ...>
    login-processing-url="/resources/j_spring_security_check"
        default-target-url="/index"
        login-page="/login" authentication-failure-url="/login?login_
logout-url="/resources/j_spring_security_logout" />
...
pattern="/login" access="permitAll" />
pattern="/**" access="isAuthenticated()" />
```

Answer:

C. In the controller function of your login page: check if a user is logged in.

then forward him to index page in that case.

Relevant code:

```

Authentication auth = SecurityContextHolder.getContext().getAuthentication();

if (!(auth instanceof AnonymousAuthenticationToken)) {

    /* The user is logged in :) */
    return new ModelAndView("forward:/index");
}

```

Q. Please rate how well the similar question and answer (section B and C) provide a solution to the first question (section A).

Very bad ☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ☐ 6 ☐ 7 ☐ 8 ☐ 9 ☐ 10
Very good

26. A. How can I convert timezone of one City to another?

I want to convert timezone of one city to another, using Joda DateTime, the process is straight forward, but I have a problem.

I am referring to Joda Database and I am unable to locate specific cities, in the database there are so few mappings say for e.g. Asia/Kolkata, if I want to search timezone for say another city like Asia/New Delhi which is in Asia or more specifically India, I cannot do that, it is mostly probable that cities within the same country have same time zone but its not always true, so I am stuck in understanding of as how to map a city to its country and vice versa.

But there is no mapping found for cities other than what is speicified in the database.

Is there any way other than or including Joda Datetime which I can use to convert by directly inputting the city name in Olson format?

I have seen many websites 1, 2 which do what I want, please help me understand what I am missing here.

Similar Question:

B. Get timezone from a country

Can anyone tel me how to get a timezone name from city and country name? Any webservice link also will suffice my need..

For eg, Input : Bangalore, India

Output is : IST

Thanks

Answer:

C. <http://www.timeanddate.com/worldclock/search.html>

Enter just the city name and it will give you all kinds of information, including the time zone.

Q. Please rate how well the similar question and answer (section B and C) provide a solution to the first question (section A).

Very bad ☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ☐ 6 ☐ 7 ☐ 8 ☐ 9 ☐ 10
Very good

27. **A. install java plugin inside HTML element in ActionScript3**

i have tried to open a link inside HTML element in AS3 , but there is nothing . in chrome,mozilla will work and open the link because it need a java plugin installed. can i add java plugin in my Adobe Air to use it in HTML element and open link ? the link is Click Here

Similar Question:

B. Can Flash SWF communicate with Java applet, and vice versa, in any way?

I know Flash can use ExternalInterface to call Javascript functions but I don't know if it can call applet also like that. Maybe it can be done by SWF -> JS -> Applet, and back, I don't know. Since I have no idea how to do applets, I would also appreciate at least some pseudo code. Thanks.

Answer:

C. Applets can communicate with JS, and JS can control applets. See these examples of Java/JavaScript interaction.

Q. Please rate how well the similar question and answer (section B and C) provide a solution to the first question (section A).

Very bad ☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ☐ 6 ☐ 7 ☐ 8 ☐ 9 ☐ 10
Very good

28. **A. Is it possible to configure a custom error message per converter?**

Is it possible to configure a custom error message per converter?

For example I have this in my xwork-conversion.properties:

```
java.util.Date=mx.com.afirme.midas2.converter.DateConverter
```

Whenever a Date conversion fails in any action I'd like to show a message like this:

Incorrect format, expected mm/dd/yyyy

I don't want to define a custom message per property as mentioned in the documentation:

However, sometimes you may wish to override this message on a per-field basis. You can do this by adding an `il8n` key associated with just your action (`Action.properties`) using the pattern `invalid.fieldvalue.xxx`, where `xxx` is the field name.

Similar Question:

B. Struts2 Customizing error message

I am using struts2 in my application and trying to display error message using `s:actionerror/`. It displays fine but a dot(.) also appears with the error message which looks ugly and is displayed like list.

Is there any way to customize the error message in struts2.

Thanks in advance.

Answer:

C. Another solution is to override the template for the default `actionError` output. You can find the default templates in the struts2 core jar. If you pull out `template.simple/actionerror.ftl`, you can customize that. You can either come up with your own template and reference it in the `s:actionerror` tag with the `template` attribute, or you can keep the same name and put it in `/template/simple` and it will be used as the default.

Most of the templates are in freemarker, although there are still some of them in velocity depending on your struts2 version. Both are pretty easy to work with templating engines.

Q. Please rate how well the similar question and answer (section B and C) provide a solution to the first question (section A).

Very bad ☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ☐ 6 ☐ 7 ☐ 8 ☐ 9 ☐ 10
Very good

29. A. JTable Customized Row

I would like some input on how to implement the following row rendering. I don't need actual code but just the concept. Row 1 is Yellow and Row 2 is White. My bean has Order Date, Req Ship Date, ... and Product as properties.

Order Date	Req Ship Date	InHand Date	Customer	Ship Date	WO#	Ship By	Ready	Shipped	Notes
Product:									
Product:									

Similar Question:**B. Change only one cell's color in JTable**

I know I can set the whole column's background color with this code, but how can I set a different color for each cell? I have a table with two columns and one to one thousand rows.

```
words.getColumn("columnNameHere").setCellRenderer(
    new DefaultTableCellRenderer() {
        @Override
        public Component getTableCellRendererComponent(JTable table, Object value,
            boolean isSelected, boolean hasFocus, int row, int column) {
            setText(value.toString());
            setBackground(Color.RED);
            return this;
        }
    }
);
```

Answer:

C. The row and column number are passed into `getTableCellRendererComponent`. So you could do something like:

```
public Component getTableCellRendererComponent(JTable table, Object value, boolean isSelected,
    boolean hasFocus, int row, int column) {
    setText(value.toString());
    if (row==12 && column==2) {
        setBackground(Color.RED);
    }
    return this;
}
```

Q. Please rate how well the similar question and answer (section B and C) provide a solution to the first question (section A).

Very bad ☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ☐ 6 ☐ 7 ☐ 8 ☐ 9 ☐ 10
Very good

30. A. local storage list implementation

Is there a implementation of List as a local storage instead of main memory resident in java. Basically my application have to store a huge amount of data and i don't want to use the memory resident list implementations.

Similar Question:**B. Are there any decent on-disk implementations of Java's Map?**

I'm looking for an on-disk implementation of java.util.Map. Nothing too fancy, just something that I can point at a directory or file and have it store its contents there, in some way it chooses. Does anyone know of such a thing?

Answer:

C. You could use a simple EHCache implementation? The nice thing about EHCache being that it can be very simple to implement :-)

I take it you've ruled out serialising / deserialising an actual Map instance?

Q. Please rate how well the similar question and answer (section B and C) provide a solution to the first question (section A).

Very bad ☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ☐ 6 ☐ 7 ☐ 8 ☐ 9 ☐ 10
Very good

31. A. Is it possible to configure a custom error message per converter?

Is it possible to configure a custom error message per converter?

For example I have this in my xwork-conversion.properties:

java.util.Date=mx.com.afirme.midas2.converter.DateConverter Whenever a Date conversion fails in any action I'd like to show a message like this:

Incorrect format, expected mm/dd/yyyy

I don't want to define a custom message per property as mentioned in the documentation:

However, sometimes you may wish to override this message on a per-field basis. You can do this by adding an i18n key associated with just your action (Action.properties) using the pattern invalid.fieldvalue.xxx, where xxx is the field name.

Similar Question:**B. remove conversion validation message in struts 2 or make it general**

In my struts2 application I have field named carrierNo that accepts integer, when i put string in it gives me this validation error message:

Invalid field value for field "carrierNo". i can customize this error message in the properties file like this

invalid.fieldvalue.carrierNo=this field does not accept characters

but i don't want to write a customized message for every non String field in my web application, i want to make it general, i tried the following but it did not work
invalid.fieldvalue.%getText(fieldName)=this field does not accept characters

if there is no way to make general, please help me disable this message at all. then i will use conversion field validator with single message that i define in the properties file.

so my request is to help me make this invalid.fieldvalue.carrierNo general something like this form invalid.fieldvalue.%getText(fieldName)

or disable the display of this error message Invalid field value for field "carrierNo".

Answer:

C. You could create your own implementation of ConversionErrorInterceptor which finds out the class of failed field and gets your custom message. Edit:

See source code for ConversionErrorInterceptor. For example you could do something like this in your custom interceptor inside intercept method

```
// get field by name from action
Field f = invocation.getAction().getClass().getDeclaredField(propertyName);
// get type of field
Class clz = f.getType();

String message = LocalizedTextUtil.findDefaultText(XWorkMessages.DEFAULT_INVALID_FIELD_VALUE_MESSAGE,
    invocationContext.getLocale());
```

And in your messages.properties file put xwork.default.invalid.fieldvalue.int, xwork.default.invalid.fieldvalue.string, etc.

Q. Please rate how well the similar question and answer (section B and C) provide a solution to the first question (section A).

Very bad ☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ☐ 6 ☐ 7 ☐ 8 ☐ 9 ☐ 10
Very good

32. A. How to create an installer out of an installer in Java EnvironMent

i must created an exe with another Exe. While the first Exe has some restrictions, in the second exe which follows that restrictions. i use Innosetup and Launch4j for creating the first exe and i don't know how to develop an exe while i am execute another exe, is it possible in java environment.

Similar Question:**B. How to create an installer out of an installer?**

I have an old legacy application around for which I only have the installer. it doesn't do anything more than uncompress and register itself and his library's.

As i don't have the source-code it gives me a lot of headaches in maintenance problems. In some particular computers (Acer Aspire One with Windows) just do not run.

I would like to extract the files and re-create this installer with NSIS. Is this possible or I'm nuts?

The original installer has been created with Ghost Installer Studio.

Answer:

C. One option is to find a machine that it does run on, and then install some "install guard" software (often called things like Acme Uninstaller). Use this to track what actually gets installed and then copy the relevant files and write your own.

It looks like a lot of hassle to me, and you may be on thin ice with regards to the licence.

Q. Please rate how well the similar question and answer (section B and C) provide a solution to the first question (section A).

Very bad ☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ☐ 6 ☐ 7 ☐ 8 ☐ 9 ☐ 10
Very good

33. A. customizing the myleadingmargin2 class to allow l right sided images

I've figured out how to wrap text around an image using the myleadingmargin2 class but it only works with the image on the left side of the text body. I know that the dir variable in drawleadingmargin can change that i just don't know when or where to set it.

Similar Question:

B.How to layout text to flow around an image

Can you please tell me if there is a way to layout text around an image? Like this:

```

----- text text text
|      | text text text
----- text text text
text text text text
text text text text

```

I have gotten a response from an android developer about this question. But I am not sure what he means by doing my own version of TextView? Thank for any tips.

On Mon, Feb 8, 2010 at 11:05 PM, Romain Guy wrote:

Hi,

This is not possible using only the supplied widgets and layouts. You could write your own version of TextView to do this, it shouldn't be hard.

Answer:

C.I can offer more comfortable constructor for The MyLeadingMarginSpan2 class

```

MyLeadingMarginSpan2(Context cc,int textSize,int height,int width) {
    int pixelsInLine=(int) (textSize*cc.getResources().getDisplayMetrics().s
    if (pixelsInLine>0 && height>0) {
        this.lines=height/pixelsInLine;
    } else {
        this.lines=0;
    }
    this.margin=width;
}

```

Q. Please rate how well the similar question and answer (section B and C) provide a solution to the first question (section A).

Very bad ☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ☐ 6 ☐ 7 ☐ 8 ☐ 9 ☐ 10
Very good

34. A. screenshot on javascript and java

I am using google visualization api for my java based website. I am planning to capture the screen shots of my clients wedpage having images generated by Google Visualization API,and i want to prepare Ppt using that screenshots.

is there any way to save these visualizations on my server so that i use them directly on my application to generate pptfile?

Suggestions please?

Similar Question:

B. Take a screenshot of a webpage with JavaScript?

Is it possible to to take a screenshot of a webpage with JavaScript and then submit that back to the server?

I'm not so concerned with browser security issues. etc. as the implementation would be for HTA. But is it possible?

Answer:

C. I have done this for an HTA by using an ActiveX control. It was pretty easy to build the control in VB6 to take the screenshot. I had to use the `keybd_event` API call because `SendKeys` can't do `PrintScreen`. Here's the code for that:

```

Declare Sub keybd_event Lib "user32" _
    (ByVal bVk As Byte, ByVal bScan As Byte, ByVal dwFlags As Long, ByVal dwExtraInfo As Long)

Public Const CaptWindow = 2

Public Sub ScreenGrab()
    keybd_event &H12, 0, 0, 0
    keybd_event &H2C, CaptWindow, 0, 0
    keybd_event &H2C, CaptWindow, &H2, 0
    keybd_event &H12, 0, &H2, 0
End Sub

```

That only gets you as far as getting the window to the clipboard.

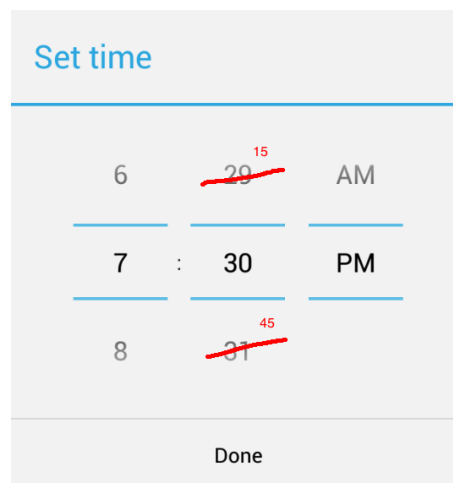
Another option, if the window you want a screenshot of is an HTA would be to just use an `XMLHttpRequest` to send the DOM nodes to the server, then create the screenshots server-side.

Q. Please rate how well the similar question and answer (section B and C) provide a solution to the first question (section A).

Very bad ☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ☐ 6 ☐ 7 ☐ 8 ☐ 9 ☐ 10
 Very good

35. A. Modify the next and previous times in a TimePickerDialog?

I've successfully implemented a TimePickerDialog that only allows the user to select a minute value that is an increment of 15, e.g. 00, 15, 30, 45. However, the "grayed out" next and previous times still show +/- 1 minute, which is unintuitive from a user perspective. Is there any way to modify the next and previous times programmatically? I have searched through the related APIs and attempted to search the view tree in a debugger, but I cannot find a way to do so.



an illustration of the unintuitive timepickerdialog

To illustrate the problem, here is the TimePickerDialog with 7:30 PM selected. When the user clicks 31, the TimePicker will actually select 7:45. How can I update the gray "next" value to show 45 instead of 31?

Similar Question:

B. Android - TimePicker minutes to 15

Can I control the Android TimePicker view to just show the minutes to 15 minutes interval?

Meaning if it's 12:28 now, show 12:30 and clicking the + and - button will increment and decrement by 15?

Thank you,

Tee

Answer:

C. Here's my version where you can set the interval:

```
private static final int TIME_PICKER_INTERVAL=15;
private boolean mIgnoreEvent=false;
```

```

private TimePicker.OnTimeChangeListener mTimePickerListener=new TimePicker.OnTimeC
    public void onTimeChanged(TimePicker timePicker, int hourOfDay, int minute){
        if (mIgnoreEvent)
            return;
        if (minute%TIME_PICKER_INTERVAL!=0){
            int minuteFloor=minute-(minute%TIME_PICKER_INTERVAL);
            minute=minuteFloor + (minute==minuteFloor+1 ? TIME_PICKER_INTERVAL : 0);
            if (minute==60)
                minute=0;
            mIgnoreEvent=true;
            timePicker.setCurrentMinute(minute);
            mIgnoreEvent=false;
        }
    }
};

```

Q. Please rate how well the similar question and answer (section B and C) provide a solution to the first question (section A).

Very bad ☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ☐ 6 ☐ 7 ☐ 8 ☐ 9 ☐ 10
Very good

36. A. Wireless networks detector on Java SE

Hi, can you advise me some materials about creating software for PC which can detect wireless networks, on Java.

Similar Question:

B. Finding SSID of a wireless network with Java We're doing a project coded in Java (compiled for JRE 1.6) and need some help with a little but apparently complicated feature: We want to do a certain action when a specific wireless network is connected e.g. when the connected SSID=="myNetworkAtHome" or similar.

After looking through this site, google and the Java documentation we have come a little closer. After looking at the code here:

<http://download.oracle.com/javase/tutorial/networking/nifs/retrieving.html>

It seems we were getting close but it hits a deadend, all the interfaces seems to be connected to "net0" through "net13" (on my laptop that is.) And we're unable to get the SSID out of any interface at all. I do realise the code in the example is

only giving the interface names and not connected networks, but it doesn't seem to offer a way of fetching the connected network information.

Any help on this would be extremely helpfull!

Regards Martin NJ

Answer:

C. You can't access this low-level details of the network in Java. You can get some details of the network interface with the `NetworkInterface` class but if you see at the provided methods, no one is related to Wifi networks nor any way to get the SSID is provided. As pointed below, you should use some native functionality through calling a native library with JNI or by calling a OS tool with Runtime.

Java is not designed to do that kind of things, is hard to implement in a platform-independent way and any hardware-level detail can not be managed in Java by principle.

Same applies to other networks like 3G, GPRS... the application should not be aware of the connection type nor its details. Java can only manage things at the Transport (TCP) level, not the network (IP) not Link (3G, Wifi, Ethernet...), so you can only manage sockets.

Q. Please rate how well the similar question and answer (section B and C) provide a solution to the first question (section A).

Very bad ☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ☐ 6 ☐ 7 ☐ 8 ☐ 9 ☐ 10
Very good

37. A. How do i pause MediaPlayer when the microphone is on?

I noticed that if I start the Google voice search my music player doesn't pause.

How do I refer to the state when the mic is on? so i can pause the mp and start it again.

Similar Question:

B. Detecting whether a headset is plugged into an Android device or not.

How can I determine whether a headset is plugged into an Android device or not?

Answer:

C. When you say "headset", do you mean "wired headset"? If so, there's an intent to detect whether or not one is being plugged or unplugged: `ACTION_HEADSET_PLUG`.

To check the status, you can use `AudioManager.isWiredHeadsetOn()`, although that may return false if there is also a bluetooth headset, and audio is routed to that instead.

Q. Please rate how well the similar question and answer (section B and C) provide a solution to the first question (section A).

Very bad ☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ☐ 6 ☐ 7 ☐ 8 ☐ 9 ☐ 10
Very good

38. **A. Make a table appear automatically in an Eclipse plugin**

How make a table appear automatically in an Eclipse plugin?

I am currently developing an eclipse plugin and I created this option in the plugin that provides me a table on the "view". However, every time when I want to see the table, I have to do this process:

Window -> Show View -> Other -> My Table

Doing this process is not user friendly. Is there a way to make this table appear automatically when I click my option button and refreshes it every time when I click it.

Similar Question:

B. Programmatically showing a View from an Eclipse Plug-in

I have a plug-in to an Eclipse RCP application that has a view. After an event occurs in the RCP application, the plug-in is instantiated, its methods are called to populate the plug-in's model, but I cannot find how to make the view appear without going to the "Show View..." menu.

I would think that there would be something in the workbench singleton that could handle this, but I have not found out how anywhere.

Answer:

C. You are probably looking for this:

```
PlatformUI.getWorkbench().getActiveWorkbenchWindow().getActivePage().showView("viewId");
```

Q. Please rate how well the similar question and answer (section B and C) provide a solution to the first question (section A).

Very bad ☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ☐ 6 ☐ 7 ☐ 8 ☐ 9 ☐ 10
Very good

39. A. How can I exclude methods from EMMA reports?

Is it possible to exclude selected methods from EMMA code coverage reports? I don't want to have getters, setters, equals methods, etc. in my final EMMA reports. I know that it is impossible to configure this in EMMA, but I'm hoping there is a trick which will allow me to skip those methods. I thought about modifying coverage.em, but I don't know the format of that file, or how to read it.

Similar Question:**B. How to force Emma code coverage report to ignore some methods?**

Some methods, such as auto-generated getters, setters, equals and toString, are trivial for test. However, if they aren't added into the testing classes, the code coverage percentage (calculated using Emma) is reduced and may crash our system build.

How can I force emma to ignore these methods in the code coverage percentage?

Regards,

Felipe

Answer:**C. From the EMMA FAQ:**

A feature to allow EMMA users to mark arbitrary methods as excluded from coverage is being considered for future versions.

So unfortunately this doesn't seem to be possible at the moment. There's an open feature request for this in EMMA's tracker.

Q. Please rate how well the similar question and answer (section B and C) provide a solution to the first question (section A).

Very bad ☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ☐ 6 ☐ 7 ☐ 8 ☐ 9 ☐ 10
Very good

40. A. Spring Roo how to maintain state between commands

I have written a couple of simple add-on's for Spring Roo. How do you maintain state between successive commands?

Similar Question:**B. maintain state with spring between requests**

I am new to spring so sorry if this is a beginners question, but the manual is not clear (at least not for me)

My question is: how do I share state between requests in spring? I can send data from the controller to the view by using a ModelMap, but the data in the ModelMap is not sent back to the next controller by the view. How can I do this with spring?

Below is a part of my source code. In the second controller the modelMap doesn't contain the data I stored in the modelMap in the first controller. How am I supposed to maintain state between controllers in spring?

thanks a lot for help.

```
@RequestMapping(value = "find/something", method = RequestMethod.GET)
public String foo(@RequestParam("parent") Parent parent, ModelMap modelMap) {
    ...
    modelMap.addAttribute("question_index", 42);
    modelMap.addAttribute("something", new Something());
    modelMap.addAttribute("data", new Data());
    return "some/view";
}

<form:form action="bla" method="POST" modelAttribute="data">
...// using Something() and 42
form:form>

@RequestMapping(value = "bla", method = RequestMethod.POST)
public String bla(@ModelAttribute("data") Data data, BindingResult result, ModelMap
System.out.println(modelMap); // doesn't contain question_index, or something
}
```

Answer:

C. You can either put the modelMap in the HttpSession, or (preferable for larger applications), use Spring web flow where you can have the so called conversations.

Q. Please rate how well the similar question and answer (section B and C) provide a solution to the first question (section A).

Very bad ☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ☐ 6 ☐ 7 ☐ 8 ☐ 9 ☐ 10
Very good

41. **A. How to generate WSDL from SOAP message sample**

I need to call some Web Services, but there's no WSDL for it. (I'm using Java)
 For now, I'm building the SOAP message manually using javax.xml.soap classes...

but that's not very clean... What I would like to know... is there a way to generate a WSDL file based on the content of a SOAP message (sample) ? Then I would be able to use CXF... Thanks.

Similar Question:

B. Generate WSDL for existing SOAP Service using captured traffic

I need to use the SOAP service of a printer. There is a windows tool to access this service and I used it to generate SOAP requests and responses of the important functions that I need.

Now I should write a client for Linux using Python and I found the easiest way would be to use the suds library with an WSDL file. But I don't have this WSDL!

As I investigated the windows tool (looked at the hexdump of the executables), I came to the conclusion that there probably is no WSDL file at all.

Now my question is, has anybody experience with "reverse engineering" SOAP services and knows tools which could be useful for creating WSDL files for existing services? (Googleing hasn't brought up anything useful yet).

Answer:

C. You mentioned this is the SOAP service of a printer. Is the printer's API documented on the manufacturer's site? Does the documentation include the WSDL? Can you get the WSDL from the manufacturer? If you can get the WSDL from the manufacturer then you're done!

If not, then you have to build the WSDL by yourself because I doubt you can find a tool that generates WSDLs given SOAP samples (when working with SOAP web services you mainly get two kinds of tools: those that generate code from WSDL + those that generate WSDL from code).

It's not hard to create the WSDL if you are familiar with SOAP, WSDL and XSD. You just need a text editor or maybe even a WSDL editor to speed things up.

If you don't have full confidence in your WSDL knowledge, there are still some tools that can get you most of the way to the complete WSDL. Here is a way you could do it:

1 - First you need to create the XML schema for the SOAP payloads. For this you can find tools, even some online. After you have the schema, tweak it to your needs by adding, changing or removing elements.

2 - Now you can use the XSD to generate a WSDL. There is an online tool that does that. It just needs the request/response element types to end with Request/Response. Make sure you read the instructions.

You take your XSD file, change the names of the operations to add the Request/Response suffix and feed it to the WSDL Generator - Web Tool. You will get your WSDL.

Now tweak this WSDL as you like (remove the Request/Response suffixes if you don't need them) then ...

3 - ... make sure you end up with a valid WSDL.

4 - Now you can take your WSDL and use a tool like SoapUI to generate sample requests and responses from it just to verify that you get the proper results back.

Do the SoapUI messages match the messages you started with? If yes, you are done and can feed the WSDL to suds to create the Linux client. If not, tweak the WSDL until you get the result you are after.

Q. Please rate how well the similar question and answer (section B and C) provide a solution to the first question (section A).

Very bad ☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ☐ 6 ☐ 7 ☐ 8 ☐ 9 ☐ 10
Very good

42. **A. Differences of pragma: no-cache and Cache-Control: no-cache**

I need to call some Web Services, but there's no WSDL for it. (I'm using Java)
 For now, I'm building the SOAP message manually using javax.xml.soap classes...
 but that's not very clean... What I would like to know... is there a way to generate
 a WSDL file based on the content of a SOAP message (sample) ? Then I would
 be able to use CXF... Thanks.

Similar Question:

B. Difference between Pragma and Cache-control headers?

I read about Pragma header on Wikipedia which says: "The Pragma: no-cache
 header field is an HTTP/1.0 header intended for use in requests. It is a means for
 the browser to tell the server and any intermediate caches that it wants a fresh
 version of the resource, not for the server to tell the browser not to cache the
 resource. Some user agents do pay attention to this header in responses, but the
 HTTP/1.1 RFC specifically warns against relying on this behavior."

But I haven't understood what it does? What is the difference between the Cache-
 Control header whose value is no-cache and Pragma whose value is also no-cache?

Answer:

C. Pragma is the HTTP/1.0 implementation and cache-control is the HTTP/1.1
 implementation of the same concept. They both are meant to prevent the client

from caching the response. Older clients may not support HTTP/1.1 which is why that header is still in use.

Q. Please rate how well the similar question and answer (section B and C) provide a solution to the first question (section A).

Very bad ☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ☐ 6 ☐ 7 ☐ 8 ☐ 9 ☐ 10
Very good

43. **A. session sharing on multiple domain but same ip address in tomcat**

I have two domains one is example.com and the second in mydomain.com both domain is hosted in same ip address. I have a feature like add to cart in the website and it is stored in session. When i store values in session in example.com and i got to the mydomain.com the session not getting in mydomain.com domain. Here the problem is domain name is different but both the domain loading the same content or same ip. On close looking i can find the jsessionid generated for the different domain are different. So my question is, is there any way to share the session between two domain? Is this issue will be fixed by making the jsessionid same for both domain? I have implemented my project in struts framework in java.

Similar Question:

B. Cookies - set across multiple domains

My company has a setup as follows:

subdomain1.domain1.com
subdomain2.domain1.com
subdomain3.domain1.com
subdomain4.domain1.com
subdomain5.domain1.com
subdomain6.domain1.com
subdomain1.domain2.com
subdomain2.domain2.com
subdomain3.domain2.com
subdomain4.domain2.com
subdomain5.domain2.com
subdomain6.domain2.com

On each site, bearing in mind there can be a hundred sites per subdomain, users can log in. We, as developers, have to test frontends across several browsers, but some work may only be required on a section once logged in.

I have written a userscript which enables us to save a username and password (and other details which I cannot mention because of confidentiality). The script checks to see if the user account exists by filling in the login form and clicking the submit button. If not, it registers for us - thus automating the registration process.

Sharing cookies between subdomains on the same domain is easy. If I am on subdomain1.domain1.com I can save a cookie which can be retrieved by subdomain2.domain1.com. However, I would also like to save these for domain2. I do not appear to be able to get this to work.

I can see two solutions from here - either:

- 1) attach an iFrame using the userscript, which loads a site on domain2. This then uses the querystring to decide what to set to what, or;
- 2) use a form with method="POST", and simply post to a file on each domain.

Either way will be resource intensive, particularly if the cookies are updated each time a cookie changes. We also have URL masking in place. So we'd also have to take into account sites like abc.clientdomain1.com, abc.clientdomain2.com etc.

Answer:

C. Create a common domain specifically for your cookies and use it as a getter/setter API.

http://cookie.domain.com/set/domain1

http://cookie.domain.com/get/domain1

http://cookie.domain.com/set/domain2

http://cookie.domain.com/get/domain2

and so on.

Q. Please rate how well the similar question and answer (section B and C) provide a solution to the first question (section A).

Very bad ☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ☐ 6 ☐ 7 ☐ 8 ☐ 9 ☐ 10
Very good

44. A. Library to manipulate JPEG in java

I'm new to java;

I need to manipulate a jpeg image to do steganography. Can you tell me if there is any library to manipulate jpeg images?

I mean is there is any function to do a DCT or quantization ??

Thanks in advance

Similar Question:**B. What is the best java image processing library/approach?**

I am using both the JAI media apis and ImageMagick?

ImageMagick has some scalability issues and the JNI based JMagick isn't attractive either. JAI has poor quality results when doing resizing operations compared to ImageMagick.

Does anyone know of any excellent tools either open source or commercial that are native java and deliver high quality results?

Answer:

C. There's ImageJ, which boasts to be the world's fastest pure Java image processing program

It can be used as a library in another application. It's architecture is not brilliant, but it does basic image processing tasks.

Q. Please rate how well the similar question and answer (section B and C) provide a solution to the first question (section A).

Very bad ☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ☐ 6 ☐ 7 ☐ 8 ☐ 9 ☐ 10
Very good

45. **A. Is there any third-party framework or technology to reload an updated(latest) jar dynamically?**

As title, I have a requirement to load the latest jar if the jar have been updated while the application is on running mode. Due to the fact that the jar status is unstable , it will be changed at any time. So guys, could you give me sort of suggestion about this? Appreciate.

The reason why jar file becomes unstable is that this project is under the programming and debugging status. And this jar need to feed for another running java application. Thus the interface might be specified but the implements are supposed to have an efficient modification. If we can dynamically reload updated jar, we can speed up the debug time.

Similar Question:**B. Can I dynamically unload and reload (other versions of the same) JAR?**

I am writing a server program which is used to run unit tests of an API (displaying lots of information and providing web access to control / monitor the whole thing)...

This API is known to the server during compile time and is provided as a JAR.

To be able to compare between unit test results of different versions of the API (without restarting the server), I want to be able to unload the 'current' version of the API, and to reload a newer one (or an older one).

I don't want to use URLClassLoader and invoke every single method by name (using `getDeclaredMethod("someMethod")`), because the server heavily depends on the API and it would be complicated to 'wrap' every method call in such dirty way.

I was thinking: Since all interfaces of all versions of the JAR are same, couldn't I do it by somehow reloading an other version of the JAR (without that by-name-invokation?).

Note: I am using latest Java SE (6) and Java EE (5).

If you think, what I'm trying to achieve is not possible, please suggest a 'workaround' or a different concept.

Answer:

C. OSGi is a framework that will allow you to do it. JSR 277 the Java Module System is designed for doing that as well (I think). I have not followed the OSGi -vs- JSR 277 debate, so I don't know if they are trying to merge them at all.

You can roll your own with class loaders, but it'll be less "fun".

Q. Please rate how well the similar question and answer (section B and C) provide a solution to the first question (section A).

Very bad ☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ☐ 6 ☐ 7 ☐ 8 ☐ 9 ☐ 10
Very good

46. A. Java equivalent of Ruby's URI.escape?

Today I had to escape URL (<http://rfc3986.openrfc.org/> - for instance, replace 'space' character with %20) and surprisingly I couldn't find any API in Java which would take unescaped URL as 1 parameter and return escaped URL.

I am aware of URI class which is capable of performing escaping operation, however it's required to use multiargument constructor

(<http://download.oracle.com/javase/1.4.2/docs/api/java/net/URI.html> #URI(java.lang.String, java.lang.String, java.lang.String, int, java.lang.String, java.lang.String, java.lang.String)), using single-argument constructor simply doesn't perform escaping but throws an exception.

Luckily, I was already emedding JRuby library so I was able to use `URI.escape` method <http://www.ruby-doc.org/stdlib/libdoc/uri/rdoc/classes/URI/Escape.html> however I'm really surprised there is no easy way to achieve same effect in Java.

Am I missing some API?

Similar Question:

B. Java - Convert String to valid URI object

I am trying to get a `java.net.URI` object from a `String`. The string has some characters which will need to be replaced by their percentage escape sequences. But when I use `URLEncoder` to encode the `String` with UTF-8 encoding, even the `/` are replaced with their escape sequences.

How can I get a valid encoded URL from a `String` object?

`http://www.google.com?q=a b` gives `http%3A%2F%2Fwww.google.com...` whereas I want the output to be `http://www.google.com?q=a%20b`

Can someone please tell me how to achieve this.

I am trying to do this in an Android app. So I have access to a limited number of libraries.

If you think, what I'm trying to achieve is not possible, please suggest a 'workaround' or a different concept.

Answer:

C. You might try: `org.apache.commons.httpclient.util.URIUtil.encodeQuery` in Apache commons-httpclient project

Like this (see `URIUtil`):

`URIUtil.encodeQuery("http://www.google.com?q=a b")` will become:

`http://www.google.com?q=a%20b`

You can of course do it yourself, but URI parsing can get pretty messy...

Q. Please rate how well the similar question and answer (section B and C) provide a solution to the first question (section A).

Very bad ☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ☐ 6 ☐ 7 ☐ 8 ☐ 9 ☐ 10
Very good

References

- Adamic, L., Zhang, J., Bakshy, E., and Ackerman, M. (2008). Knowledge sharing and yahoo answers: everyone knows something. In *Proceedings of the 17th international conference on World Wide Web*, pages 665–674. ACM.
- Adamic, L. A. and Glance, N. (2005). The political blogosphere and the 2004 us election: divided they blog. In *Proceedings of the 3rd international workshop on Link discovery*, pages 36–43. ACM.
- Agichtein, E., Castillo, C., Donato, D., Gionis, A., and Mishne, G. (2008). Finding high-quality content in social media. In *Proceedings of the international conference on Web search and web data mining*, pages 183–194. ACM.
- Airio, E., Järvelin, K., Saatsi, P., Kekäläinen, J., and Suomela, S. (2004). Ciri-an ontology-based query interface for text retrieval. In *Web Intelligence: Proceedings of the 11th Finnish Artificial Intelligence Conference*.
- Alani, H., O’Hara, K., and Shadbolt, N. (2002). Ontocopi: Methods and tools for identifying communities of practice, intelligent information processing conference. *IFIP World*.
- Albert, R., Jeong, H., and Barabási, A.-L. (1999). Internet: Diameter of the world-wide web. *Nature*, 401(6749):130–131.
- Albors, J., Ramos, J., and Hervas, J. (2008). New learning network paradigms: Communities of objectives, crowdsourcing, wikis and open source. *International Journal of Information Management*, 28(3):194 – 202.
- Alexa (2015a). Alexa: reddit.com.
- Alexa (2015b). Alexa: stackoverflow.com.
- Ambati, V., Vogel, S., and Carbonell, J. G. (2011). Towards task recommendation in micro-task markets. In *Human computation*, pages 1–4. Citeseer.
- Amitay, E., Carmel, D., Har’El, N., Ofek-Koifman, S., Soffer, A., Yogev, S., and Golbandi, N. (2009). Social search and discovery using a unified approach. In *Proceedings of the 20th ACM conference on Hypertext and hypermedia*, pages 199–208. ACM.

- Anderson, A., Huttenlocher, D., Kleinberg, J., and Leskovec, J. (2012). Discovering value from community activity on focused question answering sites: a case study of stack overflow. In *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 850–858. ACM.
- Anderson, A., Huttenlocher, D., Kleinberg, J., and Leskovec, J. (2013). Steering user behavior with badges. In *Proceedings of the 22nd international conference on World Wide Web*, pages 95–106. International World Wide Web Conferences Steering Committee.
- Andrews, D. C. (2002). Audience-specific online community design. *Communications of the ACM*, 45(4):64–68.
- Angeletou, S., Rowe, M., and Alani, H. (2011). Modelling and analysis of user behaviour in online communities. In *The Semantic Web-ISWC 2011*, pages 35–50. Springer.
- Angeletou, S., Sabou, M., and Motta, E. (2008). Semantically enriching folksonomies with flor.
- Aroyo, L., Stash, N., Wang, Y., Gorgels, P., and Rutledge, L. (2007). Chip demonstrator: Semantics-driven recommendations and museum tour generation. In *Proceedings of the 2007 International Conference on Semantic Web Challenge - Volume 295, SWC’07*, pages 17–24, Aachen, Germany, Germany. CEUR-WS.org.
- Athanasias, N., Christophides, V., and Kotzinos, D. (2004). Generating on the fly queries for the semantic web: The ics-forth graphical rql interface (grql). In *The Semantic Web-ISWC 2004*, pages 486–501. Springer.
- Auer, S., Bizer, C., Kobilarov, G., Lehmann, J., Cyganiak, R., and Ives, Z. (2007). *Dbpedia: A nucleus for a web of open data*. Springer.
- Auer, S., Dietzold, S., and Riechert, T. (2006). Ontowiki—a tool for social, semantic collaboration. In *The Semantic Web-ISWC 2006*, pages 736–749. Springer.
- Backstrom, L., Boldi, P., Rosa, M., Ugander, J., and Vigna, S. (2012). Four degrees of separation. In *Proceedings of the 4th Annual ACM Web Science Conference*, pages 33–42. ACM.
- Backstrom, L., Huttenlocher, D., Kleinberg, J., and Lan, X. (2006). Group formation in large social networks: membership, growth, and evolution. In *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 44–54. ACM.
- Bennett, J. and Lanning, S. (2007). The netflix prize. In *Proceedings of KDD cup and workshop*, volume 2007, page 35.
- Berners-Lee, T. (2003). Semantic web - architecture: Www past & future. *Semantic Web - XML2000*.

- Berners-Lee, T. (2011). Linked data-design issues (2006). URL <http://www.w3.org/DesignIssues/LinkedData.html>.
- Berners-Lee, T., Chen, Y., Chilton, L., Connolly, D., Dhanaraj, R., Hollenbach, J., Lerer, A., and Sheets, D. (2006a). Tabulator: Exploring and analyzing linked data on the @incollectionlei2006semsearch, title=Semsearch: A search engine for the semantic web, author=Lei, Yuanguai and Uren, Victoria and Motta, Enrico, booktitle=Managing Knowledge in a World of Networks, pages=238–245, year=2006, publisher=Springer semantic web. In *Proceedings of the 3rd International Semantic Web User Interaction Workshop*, volume 2006. Citeseer.
- Berners-Lee, T. et al. (2006b). The worldwideweb browser. *The World Wide Web Consortium (W3C)*. [Online]. Available: <http://www.w3.org/People/Berners-Lee/WorldWideWeb.html>, 412.
- Berners-Lee, T., Fielding, R., and Masinter, L. (2004). Uniform resource identifier (uri): Generic syntax. Technical report.
- Berners-Lee, T., Fischetti, M., and Foreword By-Dertouzos, M. L. (2000). *Weaving the Web: The original design and ultimate destiny of the World Wide Web by its inventor*. HarperInformation.
- Berners-Lee, T., Hendler, J., and Lassila, O. (2001). The semantic web: Scientific american. *Scientific American*, 284(5):34–43.
- Berners-Lee, T. J. (2007). Giant global graph. *timbl's blog*.
- Bernstein, M. S., Monroy-Hernández, A., Harry, D., André, P., Panovich, K., and Vargas, G. G. (2011). 4chan and/b: An analysis of anonymity and ephemerality in a large online community. In *ICWSM*.
- Bharat, K. and Henzinger, M. R. (1998). Improved algorithms for topic distillation in a hyperlinked environment. In *Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval*, pages 104–111. ACM.
- Bhogal, J., Macfarlane, A., and Smith, P. (2007). A review of ontology based query expansion. *Information processing & management*, 43(4):866–886.
- Bian, J., Liu, Y., Agichtein, E., and Zha, H. (2008). Finding the right facts in the crowd: factoid question answering over social media. In *Proceedings of the 17th international conference on World Wide Web*, pages 467–476. ACM.
- Bishop, J. (2007). Increasing participation in online communities: A framework for human–computer interaction. *Computers in human behavior*, 23(4):1881–1893.
- Bizer, C., Heath, T., and Berners-Lee, T. (2009). Linked data-the story so far. *Semantic Services, Interoperability and Web Applications: Emerging Concepts*, pages 205–227.

- Board, A. R. (2007). Rss 2.0 specification.
- Bobadilla, J., Ortega, F., Hernando, A., and Gutiérrez, A. (2013). Recommender systems survey. *Knowledge-Based Systems*, 46:109–132.
- Bojars, U., Breslin, J., Peristeras, V., Tummarello, G., and Decker, S. (2008a). Interlinking the social web with semantics. *Intelligent Systems, IEEE*, 23(3):29–40.
- Bojars, U., Passant, A., Cyganiak, R., and Breslin, J. (2008b). Weaving sioc into the web of linked data. In *Linked Data on the Web (LDOW2008)*.
- Brabham, D. C. (2008). Crowdsourcing as a model for problem solving an introduction and cases. *Convergence: the international journal of research into new media technologies*, 14(1):75–90.
- Brauer, F., Barczynski, W., Hackenbroich, G., Schramm, M., Mocan, A., and Förster, F. (2009). Rankie: document retrieval on ranked entity graphs. *Proceedings of the VLDB Endowment*, 2(2):1578–1581.
- Bray, T., Paoli, J., Sperberg-McQueen, C. M., Maler, E., and Yergeau, F. (1998). Extensible markup language (xml). *World Wide Web Consortium Recommendation REC-xml-19980210*. <http://www.w3.org/TR/1998/REC-xml-19980210>, 16.
- Breslin, J. and Decker, S. (2007). The future of social networks on the internet: The need for semantics. *Internet Computing, IEEE*, 11(6):86–90.
- Breslin, J., Harth, A., Bojars, U., and Decker, S. (2005). Towards semantically-interlinked online communities. *The Semantic Web: Research and Applications*, pages 71–83.
- Brickley, D. and Guha, R. V. (2000). Resource description framework (rdf) schema specification 1.0: W3c candidate recommendation 27 march 2000.
- Brickley, D. and Miller, L. (2010). Foaf vocabulary specification 0.98. *Namespace Document*, 9.
- Bu, J., Tan, S., Chen, C., Wang, C., Wu, H., Zhang, L., and He, X. (2010). Music recommendation by unified hypergraph: combining social media information and music content. In *Proceedings of the international conference on Multimedia*, pages 391–400. ACM.
- Buhrmester, M., Kwang, T., and Gosling, S. D. (2011). Amazon’s mechanical turk a new source of inexpensive, yet high-quality, data? *Perspectives on psychological science*, 6(1):3–5.
- Bunescu, R. C. and Pasca, M. (2006). Using encyclopedic knowledge for named entity disambiguation. In *EACL*, volume 6, pages 9–16.

- Buscaldi, D., Rosso, P., and Arnal, E. S. (2005). A wordnet-based query expansion method for geographical information retrieval. In *Working notes for the CLEF workshop*.
- carmichael561 (2013). Harry speaks python.
- Castells, P., Fernandez, M., and Vallet, D. (2007). An adaptation of the vector-space model for ontology-based information retrieval. *Knowledge and Data Engineering, IEEE Transactions on*, 19(2):261–272.
- Chakrabarti, S., Dom, B. E., Kumar, S. R., Raghavan, P., Rajagopalan, S., Tomkins, A., Gibson, D., and Kleinberg, J. (1999). Mining the web’s link structure. *Computer*, 32(8):60–67.
- Chi, E. (2009). Augmented social cognition: using social web technology to enhance the ability of groups to remember, think, and reason. In *Proceedings of the 35th SIGMOD international Conference on Management of Data*, pages 973–984. ACM.
- Chi, Y., Tseng, B. L., and Tatemura, J. (2006). Eigen-trend: trend analysis in the blogosphere based on singular value decompositions. In *Proceedings of the 15th ACM international conference on Information and knowledge management*, pages 68–77. ACM.
- Chi, Y., Zhu, S., Song, X., Tatemura, J., and Tseng, B. L. (2007). Structural and temporal analysis of the blogosphere through community factorization. In *Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 163–172. ACM.
- Chin, A. and Chignell, M. (2006). A social hypertext model for finding community in blogs. In *Proceedings of the seventeenth conference on Hypertext and hypermedia*, pages 11–22. ACM.
- CIA, U. (2010). The world factbook. *Retrieved August, 20:2010*.
- Conner-Simons, A. (2015). Web inventor tim berners-lee’s next project: A platform that gives users control of their data. *CSAIL News*.
- Cooley, C. H. (1992). *Human nature and the social order*. Transaction Publishers.
- Corlosquet, S., Delbru, R., Clark, T., Polleres, A., and Decker, S. (2009). *Produce and Consume Linked Data with Drupal!* Springer.
- Correndo, G. and Alani, H. (2007). Survey of tools for collaborative knowledge construction and sharing. In *Web Intelligence and Intelligent Agent Technology Workshops, 2007 IEEE/WIC/ACM International Conferences on*, pages 7–10. IEEE.
- Cucerzan, S. (2007). Large-scale named entity disambiguation based on wikipedia data. In *EMNLP-CoNLL*, volume 7, pages 708–716.

- Cyganiak, R. (2005). A relational algebra for sparql. *Digital Media Systems Laboratory HP Laboratories Bristol. HPL-2005-170*, page 35.
- Daoud, M., Tamine, L., and Boughanem, M. (2010). A personalized graph-based document ranking model using a semantic user profile. In *User Modeling, Adaptation, and Personalization*, pages 171–182. Springer.
- Davidson, J., Liebal, B., Liu, J., Nandy, P., Van Vleet, T., Gargi, U., Gupta, S., He, Y., Lambert, M., Livingston, B., et al. (2010). The youtube video recommendation system. In *Proceedings of the fourth ACM conference on Recommender systems*, pages 293–296. ACM.
- Davoodi, E., Kianmehr, K., and Afsharchi, M. (2013). A semantic social network-based expert recommender system. *Applied intelligence*, 39(1):1–13.
- Deerwester, S. C., Dumais, S. T., Landauer, T. K., Furnas, G. W., and Harshman, R. A. (1990). Indexing by latent semantic analysis. *JAsIs*, 41(6):391–407.
- Dill, S., Eiron, N., Gibson, D., Gruhl, D., Guha, R., Jhingran, A., Kanungo, T., Rajagopalan, S., Tomkins, A., Tomlin, J. A., et al. (2003). Sementag and seeker: Bootstrapping the semantic web via automated semantic annotation. In *Proceedings of the 12th international conference on World Wide Web*, pages 178–186. ACM.
- Ding, L., Finin, T., Joshi, A., Pan, R., Cost, R. S., Peng, Y., Reddivari, P., Doshi, V., and Sachs, J. (2004). Swoogle: a search and metadata engine for the semantic web. In *Proceedings of the thirteenth ACM international conference on Information and knowledge management*, pages 652–659. ACM.
- Eddy, S. R. (1998). Profile hidden markov models. *Bioinformatics*, 14(9):755–763.
- Eddy, S. R. et al. (2009). A new generation of homology search tools based on probabilistic inference. In *Genome Inform*, number 1, pages 205–211. World Scientific.
- Egozi, O., Markovitch, S., and Gabrilovich, E. (2011). Concept-based information retrieval using explicit semantic analysis. *ACM Transactions on Information Systems (TOIS)*, 29(2):8.
- Eiron, N. and McCurley, K. S. (2003). Analysis of anchor text for web search. In *Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval*. ACM.
- El-Korany, A. (2013). Integrated expert recommendation model for online communities. *arXiv preprint arXiv:1311.3394*.
- Ellison, N. B., Steinfield, C., and Lampe, C. (2007). The benefits of facebook “friends:” social capital and college students’ use of online social network sites. *Journal of Computer-Mediated Communication*, 12(4):1143–1168.

- Ereteo, G., Buffa, M., Gandon, F., and Corby, O. (2009). *Analysis of a real online social network using semantic web frameworks*. Springer.
- Estellés-Arolas, E. and González-Ladrón-de Guevara, F. (2012). Towards an integrated crowdsourcing definition. *Journal of Information science*, 38(2):189–200.
- Evans, B. and Chi, E. (2008). Towards a model of understanding social search. In *Proceedings of the 2008 ACM conference on Computer supported cooperative work*, pages 485–494. ACM.
- Facebook (2015). Facebook company information- stats 2015. *Facebook Company Information- Stats*.
- Fazel-Zarandi, M., Devlin, H. J., Huang, Y., and Contractor, N. (2011). Expert recommendation based on social drivers, social network analysis, and semantic data representation. In *Proceedings of the 2nd International Workshop on Information Heterogeneity and Fusion in Recommender Systems*, pages 41–48. ACM.
- Felfernig, A., Jeran, M., Ninaus, G., Reinfrank, F., and Reiterer, S. (2013). Toward the next generation of recommender systems: applications and research challenges. In *Multimedia services in intelligent environments*, pages 81–98. Springer.
- Fellbaum, C. (1998). *WordNet*. Wiley Online Library.
- Fernández, M., Cantador, I., López, V., Vallet, D., Castells, P., and Motta, E. (2011). Semantically enhanced information retrieval: an ontology-based approach. *Web Semantics: Science, Services and Agents on the World Wide Web*, 9(4):434–452.
- Ferré, S. and Hermann, A. (2011). Semantic search: Reconciling expressive querying and exploratory search. In *The Semantic Web-ISWC 2011*, pages 177–192. Springer.
- Fikes, R., Hayes, P., and Horrocks, I. (2004). Owl-ql—a language for deductive query answering on the semantic web. *Web semantics: Science, services and agents on the World Wide Web*, 2(1):19–29.
- Fitzpatrick, B. and Recordon, D. (2007). Thoughts on the social graph. *Brad Fitzpatrick's Blog*, 17:52.
- Flake, G., Lawrence, S., Giles, C., and Coetzee, F. (2002). Self-organization and identification of web communities. *Computer*, 35(3):66–70.
- Freitas, A., Oliveira, J. G., O’riain, S., Da Silva, J. C., and Curry, E. (2013). Querying linked data graphs using semantic relatedness: A vocabulary independent approach. *Data & Knowledge Engineering*, 88:126–141.
- Garcia, A., Szomszor, M., Alani, H., and Corcho, O. (2009). Preliminary results in tag disambiguation using dbpedia. *First International Workshop on Collective Knowledge Capturing and Representation*.

- García-Crespo, Á., López-Cuadrado, J. L., Colomo-Palacios, R., González-Carrasco, I., and Ruiz-Mezcua, B. (2011). Sem-fit: A semantic based expert system to provide recommendations in the tourism domain. *Expert systems with applications*, 38(10):13310–13319.
- Garfield, E. et al. (1972). Citation analysis as a tool in journal evaluation. American Association for the Advancement of Science.
- Garton, L., Haythornthwaite, C., and Wellman, B. (1997). Studying online social networks. *Journal of Computer-Mediated Communication*, 3(1):0–0.
- Ghosh, A. and Hummel, P. (2011). A game-theoretic analysis of rank-order mechanisms for user-generated content. In *12th ACM Conference on Electronic Commerce (EC)*.
- Ghosh, A. and McAfee, P. (2011). Incentivizing high-quality user-generated content. In *Proceedings of the 20th international conference on World wide web*, pages 137–146. ACM.
- Glaser, H., Jaffri, A., and Millard, I. (2009). Managing co-reference on the semantic web. In *WWW2009 Workshop: Linked Data on the Web (LDOW2009)*.
- Glaser, H., Lewy, T., Millard, I., and Dowling, B. (2007). On coreference and the semantic web. Technical report, University of Southampton.
- Glaser, H., Millard, I., and Jaffri, A. (2008). Rkbexplorer.com: a knowledge driven infrastructure for linked data providers. In *European Semantic Web Conference*, volume 5021/2, pages 797–801. Springer. Event Dates: 1-5 June 2008.
- Glimm, C. O. B. (2012). {SPARQL 1.1 Entailment Regimes}.
- Grootjen, F. A. and Van Der Weide, T. P. (2006). Conceptual query expansion. *Data & Knowledge Engineering*, 56(2):174–193.
- Gruber, T. (2008). Collective knowledge systems: Where the social web meets the semantic web. *Web Semantics: Science, Services and Agents on the World Wide Web*, 6(1):4 – 13. Semantic Web and Web 2.0.
- Guha, R., McCool, R., and Miller, E. (2003). Semantic search. In *Proceedings of the 12th international conference on World Wide Web*, pages 700–709. ACM.
- Guy (2008). What’s the difference between javascript and java?
- Han, X. and Zhao, J. (2010). Structural semantic relatedness: a knowledge-based method to named entity disambiguation. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 50–59. Association for Computational Linguistics.

- Hao, T., Lu, Z., Wang, S., Zou, T., Gu, S., and Wenyin, L. (2008). Categorizing and ranking search engine's results by semantic similarity. In *Proceedings of the 2nd international conference on Ubiquitous information management and communication*, pages 284–288. ACM.
- Hartig, O. and Heese, R. (2007). The sparql query graph model for query optimization. In *The Semantic Web: Research and Applications*, pages 564–578. Springer.
- Harvey, R. (2011). Strangest language feature.
- Hawke, S., Herman, I., Archer, P., and Prud'hommeaux, E. (2015). W3c semantic web activity.
- Heath, T. and Bizer, C. (2011). Linked data: Evolving the web into a global data space. *Synthesis lectures on the semantic web: theory and technology*, 1(1):1–136.
- Heflin, J. and Hendler, J. (2000). *Searching the Web with SHOE*. Defense Technical Information Center.
- Hendler, J. (2009). Web 3.0 emerging. *Computer*, 42(1):111–113.
- Herring, S. C., Kouper, I., Paolillo, J. C., Scheidt, L. A., Tyworth, M., Welsch, P., Wright, E., and Yu, N. (2005). Conversations in the blogosphere: an analysis” from the bottom up”. In *System Sciences, 2005. HICSS'05. Proceedings of the 38th Annual Hawaii International Conference on*, pages 107b–107b. IEEE.
- Hogan, A., Harth, A., Umbrich, J., Kinsella, S., Polleres, A., and Decker, S. (2011). Searching and browsing linked data with swse: The semantic web search engine. *Web semantics: science, services and agents on the world wide web*, 9(4):365–401.
- Howe, J. (2006). The rise of crowdsourcing. *Wired magazine*, 14(6):1–4.
- Huberman, B. A., Romero, D. M., and Wu, F. (2008). Social networks that matter: Twitter under the microscope. *Available at SSRN 1313405*.
- Huffaker, D. (2010). Dimensions of leadership and social influence in online communities. *Human Communication Research*, 36(4):593–617.
- Inc, C. (2015). Stardog 4: The manual.
- ITU (2015). Ict facts and figures – the world in 2015. *ICT Facts and Figures*.
- Jain, S., Chen, Y., and Parkes, D. (2009). Designing incentives for online question and answer forums. In *Proceedings of the 10th ACM conference on Electronic commerce*, pages 129–138. ACM.
- Jain, S. and Parkes, D. (2009). The role of game theory in human computation systems. In *Proceedings of the ACM SIGKDD Workshop on Human Computation*, pages 58–61. ACM.

- Jamali, M. and Abolhassani, H. (2006). Different aspects of social network analysis. In *Web Intelligence, 2006. WI 2006. IEEE/WIC/ACM International Conference on*, pages 66–72. IEEE.
- Jeff (2012). Regex match open tags except xhtml self-contained tags.
- Jiang, J. J. and Conrath, D. W. (1997). Semantic similarity based on corpus statistics and lexical taxonomy. *arXiv preprint cmp-lg/9709008*.
- Joinson, A. N. (2008). Looking at, looking up or keeping up with people?: motives and use of facebook. In *Proceedings of the SIGCHI conference on Human Factors in Computing Systems*, pages 1027–1036. ACM.
- Kandogan, E., Krishnamurthy, R., Raghavan, S., Vaithyanathan, S., and Zhu, H. (2006). Avatar semantic search: a database approach to information retrieval. In *Proceedings of the 2006 ACM SIGMOD international conference on Management of data*, pages 790–792. ACM.
- Kane, G. C. (2009). It’s a network, not an encyclopedia: A social network perspective on wikipedia collaboration. In *Academy of Management Proceedings*, volume 2009, pages 1–6. Academy of Management.
- Kara, S., Alan, Ö., Sabuncu, O., Akpınar, S., Cicekli, N. K., and Alpaslan, F. N. (2012). An ontology-based retrieval system using semantic indexing. *Information Systems*, 37(4):294–305.
- Katz, N., Lazer, D., Arrow, H., and Contractor, N. (2004). Network theory and small groups. *Small group research*, 35(3):307–332.
- Khatib, F., DiMaio, F., Cooper, S., Kazmierczyk, M., Gilski, M., Krzywda, S., Zabranska, H., Pichova, I., Thompson, J., Popović, Z., et al. (2011). Crystal structure of a monomeric retroviral protease solved by protein folding game players. *Nature structural & molecular biology*, 18(10):1175–1177.
- Kifer, M. and Boley, H. (2010). Rif overview. *W3C Working Group Note*.
- Kim, H. L., Yang, S.-K., Song, S.-J., Breslin, J. G., and Kim, H.-G. (2007). Tag mediated society with scot ontology. In *Semantic Web Challenge*.
- Kittur, A., Chi, E., Pendleton, B. A., Suh, B., and Mytkowicz, T. (2007). Power of the few vs. wisdom of the crowd: Wikipedia and the rise of the bourgeoisie. *World wide web*, 1(2):19.
- Kleinberg, J. (1999). Authoritative sources in a hyperlinked environment. *Journal of the ACM (JACM)*, 46(5):604–632.
- Klyne, G. and Carroll, J. J. (2006). Resource description framework (rdf): Concepts and abstract syntax.

- Kogalovsky, M. R. (2012). Ontology-based data access systems. *Programming and Computer Software*, 38(4):167–182.
- Kogut, B. (2000). The network as knowledge: Generative rules and the emergence of structure. *Strategic management journal*, 21(3):405–425.
- Kosala, R. and Blockeel, H. (2000). Web mining research: A survey. *ACM Sigkdd Explorations Newsletter*, 2(1):1–15.
- Koutsomitropoulos, D. A., Domenech, R. B., and Solomou, G. D. (2011). A structured semantic query interface for reasoning-based search and retrieval. In *The Semantic Web: Research and Applications*, pages 17–31. Springer.
- Krauss (2014). Semantic web stack, 2014.
- Kwak, H., Lee, C., Park, H., and Moon, S. (2010). What is twitter, a social network or a news media? In *Proceedings of the 19th international conference on World wide web*, pages 591–600. ACM.
- Lamberti, F., Sanna, A., and Demartini, C. (2009). A relation-based page rank algorithm for semantic web search engines. *Knowledge and Data Engineering, IEEE Transactions on*, 21(1):123–136.
- Lampe, C., Wash, R., Velasquez, A., and Ozkaya, E. (2010). Motivations to participate in online communities. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 1927–1936. ACM.
- Langegger, A., Wöß, W., and Blöchl, M. (2008). *A semantic web middleware for virtual data integration on the web*. Springer.
- Langley, P. and Simon, H. A. (1995). Applications of machine learning and rule induction. *Communications of the ACM*, 38(11):54–64.
- Lawrence, S., Giles, L. C., and Bollacker, K. (1999). Digital libraries and autonomous citation indexing. *Computer*, 32(6):67–71.
- Lawrence Page, Sergey Brin, R. M. and Winograd, T. (1999). The pagerank citation ranking: Bringing order to the web. Technical report, Stanford InfoLab.
- Lazar, J. and Preece, J. (1998). Classification schema for online communities. *AMCIS 1998 Proceedings*, page 30.
- Lei, Y., Uren, V., and Motta, E. (2006). Semsearch: A search engine for the semantic web. In *Managing Knowledge in a World of Networks*, pages 238–245. Springer.
- Lévy, P. (1997). *Collective intelligence*. Plenum/Harper Collins.

- Li, J., Yang, J.-J., Liu, C., Zhao, Y., Liu, B., and Shi, Y. (2014). Exploiting semantic linkages among multiple sources for semantic information retrieval. *Enterprise Information Systems*, 8(4):464–489.
- Li, J.-Q., Zhao, Y., and Garcia-Molina, H. (2012). A path-based approach for web page retrieval. *World Wide Web*, 15(3):257–283.
- Li, Y., Bandar, Z., McLean, D., et al. (2003). An approach for measuring semantic similarity between words using multiple information sources. *Knowledge and Data Engineering, IEEE Transactions on*, 15(4):871–882.
- Li, Y., Wang, Y., and Huang, X. (2007). A relation-based search engine in semantic web. *Knowledge and Data Engineering, IEEE Transactions on*, 19(2):273–282.
- Lin, H.-F. and Lee, G.-G. (2006). Determinants of success for online communities: an empirical study. *Behaviour & Information Technology*, 25(6):479–488.
- Lin, N., Cook, K. S., and Burt, R. S. (2001). *Social capital: Theory and research*. Transaction Publishers.
- Linden, G., Smith, B., and York, J. (2003). Amazon. com recommendations: Item-to-item collaborative filtering. *Internet Computing, IEEE*, 7(1):76–80.
- Lintott, C. J., Schawinski, K., Slosar, A., Land, K., Bamford, S., Thomas, D., Raddick, M. J., Nichol, R. C., Szalay, A., Andreescu, D., et al. (2008). Galaxy zoo: morphologies derived from visual inspection of galaxies from the sloan digital sky survey. *Monthly Notices of the Royal Astronomical Society*, 389(3):1179–1189.
- Liu, S., Liu, F., Yu, C., and Meng, W. (2004). An effective approach to document retrieval via utilizing wordnet and recognizing phrases. In *Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 266–272. ACM.
- Lopez, V., Uren, V., Motta, E., and Pasin, M. (2007). Aqualog: An ontology-driven question answering system for organizational semantic intranets. *Web Semantics: Science, Services and Agents on the World Wide Web*, 5(2):72–105.
- Ludford, P. J., Cosley, D., Frankowski, D., and Terveen, L. (2004). Think different: increasing online community participation using uniqueness and group dissimilarity. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 631–638. ACM.
- Maedche, A., Staab, S., Stojanovic, N., Studer, R., and Sure, Y. (2003). Semantic portal-the seal approach. *Spinning the Semantic Web*, pages 317–359.
- Maisonneuve, N., Stevens, M., Niessen, M. E., and Steels, L. (2009). Noisetube: Measuring and mapping noise pollution with mobile phones. In *Information Technologies in Environmental Engineering*, pages 215–228. Springer.

- Mäkelä, E. (2005). Survey of semantic search research. In *Proceedings of the seminar on knowledge management on the semantic web*. Department of Computer Science, University of Helsinki, Helsinki.
- Mangold, C. (2007). A survey and classification of semantic search approaches. *International Journal of Metadata, Semantics and Ontologies*, 2(1):23–34.
- Marathe, J. (1999). Creating community online. *Durlacher Research Ltd*, pages 281–300.
- Matthews, B. (2005). Semantic web technologies. *E-learning*, 6(6):8.
- McBride, B. (2002). Jena: A semantic web toolkit. *IEEE Internet computing*, (6):55–59.
- McGuinness, D. L., Van Harmelen, F., et al. (2004). Owl web ontology language overview. *W3C recommendation*, 10(10):2004.
- McPherson, M., Smith-Lovin, L., and Cook, J. M. (2001). Birds of a feather: Homophily in social networks. *Annual review of sociology*, pages 415–444.
- Mendes, P. N., Jakob, M., García-Silva, A., and Bizer, C. (2011). Dbpedia spotlight: shedding light on the web of documents. In *Proceedings of the 7th International Conference on Semantic Systems*, pages 1–8. ACM.
- Mika, P. (2005). Flink: Semantic web technology for the extraction and analysis of social networks. *Web Semantics: Science, Services and Agents on the World Wide Web*, 3(2):211–223.
- Milgram, S. (1967). The small world problem. *Psychology today*, 2(1):60–67.
- Millard, I., Glaser, H., Salvadores, M., and Shadbolt, N. (2010). Consuming multiple linked data sources: Challenges and Experiences. Technical report, University of Southampton.
- Milne, D. and Witten, I. (2012). An open-source toolkit for mining wikipedia. *Artificial Intelligence*.
- Moldovan, D. I. and Mihalcea, R. (2000). Using wordnet and lexical operators to improve internet searches. *IEEE Internet Computing*, (1):34–43.
- Monge, P. and Contractor, N. (2003). *Theories of communication networks*. Oxford University Press, USA.
- Mäkelä, E., Hyvönen, E., Saarela, S., and Viljanen, K. (2004). Ontoviews – a tool for creating semantic web portals. In McIlraith, S., Plexousakis, D., and van Harmelen, F., editors, *The Semantic Web – ISWC 2004*, volume 3298 of *Lecture Notes in Computer Science*, pages 797–811. Springer Berlin Heidelberg.
- Nebhi, K. (2013). Named entity disambiguation using freebase and syntactic parsing.

- Newman, M. (2004). Detecting community structure in networks. *The European Physical Journal B-Condensed Matter and Complex Systems*, 38(2):321–330.
- Nielsen and McKinsey (2011). Nm incite social media report. Technical report, NM Incite.
- Nonnecke, B., Andrews, D., and Preece, J. (2006). Non-public and public online community participation: Needs, attitudes and behavior. *Electronic Commerce Research*, 6(1):7–20.
- O’reilly, T. (2007). What is web 2.0: Design patterns and business models for the next generation of software. *Communications & strategies*, (1):17.
- Paolacci, G., Chandler, J., and Ipeirotis, P. G. (2010). Running experiments on amazon mechanical turk. *Judgment and Decision making*, 5(5):411–419.
- Park, D. H., Kim, H. K., Choi, I. Y., and Kim, J. K. (2012). A literature review and classification of recommender systems research. *Expert Systems with Applications*, 39(11):10059–10072.
- Passant, A. and Laublet, P. (2008). Meaning of a tag: A collaborative approach to bridge the gap between tagging and linked data. In *LDOW*.
- Plant, R. (2004). Online communities. *Technology in society*, 26(1):51–65.
- Preece, J. (2000). *Online communities: Designing usability and supporting socialbilty*. John Wiley & Sons, Inc.
- Preece, J., Nonnecke, B., and Andrews, D. (2004). The top five reasons for lurking: improving community experiences for everyone. *Computers in human behavior*, 20(2):201–223.
- programmer564698 (2015). My first year and a half as a professional software developer: what i expected and what i got.
- Prud’Hommeaux, E., Seaborne, A., et al. (2008). Sparql query language for rdf. *W3C recommendation*, 15.
- Prud’hommeaux, E. (2004). Optimal rdf access to relational databases.
- Quilitz, B. and Leser, U. (2008). *Querying distributed RDF data sources with SPARQL*. Springer.
- Raban, D. R., Moldovan, M., and Jones, Q. (2010). An empirical study of critical mass and online community survival. In *Proceedings of the 2010 ACM conference on Computer supported cooperative work*, pages 71–80. ACM.
- Reuters, T. (2008). Opencalais. Retrieved June, 16.

- Richardson, M. and Domingos, P. (2003). Building large knowledge bases by mass collaboration. In *Proceedings of the 2nd international conference on Knowledge capture*, pages 129–137. ACM.
- Ridings, C. and Gefen, D. (2004). Virtual community attraction: Why people hang out online. *Journal of Computer-Mediated Communication*, 10(1):00–00.
- Rinaldi, A. M. (2009). An ontology-driven approach for semantic information retrieval on the web. *ACM Transactions on Internet Technology*, 9(3):10.
- Rocha, C., Schwabe, D., and Aragao, M. P. (2004). A hybrid approach for searching in the semantic web. In *Proceedings of the 13th international conference on World Wide Web*, pages 374–383. ACM.
- Rohloff, K., Dean, M., Emmons, I., Ryder, D., and Sumner, J. (2007). An evaluation of triple-store technologies for large data stores. In *On the Move to Meaningful Internet Systems 2007: OTM 2007 Workshops*, pages 1105–1114. Springer.
- Rowe, M. et al. (2009). Interlinking distributed social graphs. In *Proc. Linked Data on the Web Workshop, WWW09. Madrid, Spain*.
- Salton, G., Wong, A., and Yang, C.-S. (1975). A vector space model for automatic indexing. *Communications of the ACM*, 18(11):613–620.
- Sauermann, L. and Cyganiak, R. (2008). Cool uris for the semantic web. *W3C Interest Group Note*.
- Schmachtenberg, M., Bizer, C., Jentzsch, A., and Cyganiak, R. (2014). Linking open data cloud diagram, 2014.
- Schmidt, M., Meier, M., and Lausen, G. (2010). Foundations of sparql query optimization. In *Proceedings of the 13th International Conference on Database Theory*, pages 4–33. ACM.
- Schütze, H. (1998). Automatic word sense discrimination. *Computational linguistics*, 24(1):97–123.
- Sebastiani, F. (2002). Machine learning in automated text categorization. *ACM computing surveys (CSUR)*, 34(1):1–47.
- Shadbolt, N., Hall, W., and Berners-Lee, T. (2006). The semantic web revisited. *Intelligent Systems, IEEE*, 21(3):96–101.
- Shadbolt, N. R., Gibbins, N., Glaser, H., Harris, S., and m.c. schraefel (2004). Walking through {CS} {AKTive} space: a demonstration of an integrated semantic web application. *Web Semantics: Science, Services and Agents on the World Wide Web*, 1(4):415 – 419. International Semantic Web Conference 2003.

- Shang, S., Hui, P., Kulkarni, S. R., and Cuff, P. W. (2011). Wisdom of the crowd: Incorporating social influence in recommendation models. In *Parallel and Distributed Systems (ICPADS), 2011 IEEE 17th International Conference on*, pages 835–840. IEEE.
- Silvertown, J. (2009). A new dawn for citizen science. *Trends in ecology & evolution*, 24(9):467–471.
- Sinclair, J. and Cardew-Hall, M. (2008). The folksonomy tag cloud: when is it useful? *Journal of Information Science*, 34(1):15–29.
- Singh, P. and Shadbolt, N. (2013). Linked data in crowdsourcing purposive social network. In *Proceedings of the 22nd international conference on World Wide Web companion*, pages 913–918. International World Wide Web Conferences Steering Committee.
- Singh, V., Jain, R., and Kankanhalli, M. (2009). Motivating contributors in social media networks. In *Proceedings of the first SIGMM workshop on Social media*, pages 11–18. ACM.
- Singhal, A. (2012). Introducing the knowledge graph: things, not strings. *Official Google Blog*, May.
- Sirin, E., Parsia, B., Grau, B. C., Kalyanpur, A., and Katz, Y. (2007). Pellet: A practical owl-dl reasoner. *Web Semantics: science, services and agents on the World Wide Web*, 5(2):51–53.
- Skeet, J. (2009). What’s your most controversial programming opinion?
- Smith, E. (2008). Social relationships and groups: New insights on embodied and distributed cognition. *Cognitive Systems Research*, 9(1):24–32.
- StackOverflow (2008). Welcome to stackoverflow.
- Stankovic, M., Wagner, C., Jovanovic, J., and Laublet, P. (2010). Looking for experts? what can linked data do for you? In *LDOW*.
- Stenmark, D. (2002). Information vs. knowledge: The role of intranets in knowledge management. In *System Sciences, 2002. HICSS. Proceedings of the 35th Annual Hawaii International Conference on*, pages 928–937. IEEE.
- Stocker, M., Seaborne, A., Bernstein, A., Kiefer, C., and Reynolds, D. (2008). Sparql basic graph pattern optimization using selectivity estimation. In *Proceedings of the 17th international conference on World Wide Web*, pages 595–604. ACM.
- Surowiecki, J. (2005). *The wisdom of crowds*. Anchor.
- Tang, J., Zhang, J., Zhang, D., Yao, L., Zhu, C., Li, J.-Z., et al. (2007). Arnetminer: An expertise oriented search system for web community. In *Semantic Web Challenge*.

- Tantipathananandh, C., Berger-Wolf, T., and Kempe, D. (2007). A framework for community identification in dynamic social networks. In *Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 717–726. ACM.
- Tran, T., Cimiano, P., Rudolph, S., and Studer, R. (2007). *Ontology-based interpretation of keywords for semantic search*. Springer.
- Treude, C., Barzilay, O., and Storey, M. (2011). How do programmers ask and answer questions on the web?: Nier track. In *Software Engineering (ICSE), 2011 33rd International Conference on*, pages 804–807. IEEE.
- Trillo, R., Gracia, J., Espinoza, M., and Mena, E. (2007). Discovering the semantics of user keywords. *J. UCS*, 13(12):1908–1935.
- Tsialiamanis, P., Sidiourgos, L., Fundulaki, I., Christophides, V., and Boncz, P. (2012). Heuristics-based query optimisation for sparql. In *Proceedings of the 15th International Conference on Extending Database Technology*, pages 324–335. ACM.
- Van Damme, C., Hepp, M., and Siorpaes, K. (2007). Folksontology: An integrated approach for turning folksonomies into ontologies. *Bridging the Gap between Semantic Web and Web*, 2(2):57–70.
- Vandic, D., Van Dam, J.-W., and Frasinicar, F. (2012). Faceted product search powered by the semantic web. *Decision Support Systems*, 53(3):425–437.
- Varelas, G., Voutsakis, E., Raftopoulou, P., Petrakis, E. G., and Milios, E. E. (2005). Semantic similarity methods in wordnet and their application to information retrieval on the web. In *Proceedings of the 7th annual ACM international workshop on Web information and data management*, pages 10–16. ACM.
- Vassileva, J. (2012). Motivating participation in social computing applications: a user modeling perspective. *User Modeling and User-Adapted Interaction*, 22(1-2):177–201.
- Venkataramani, R., Gupta, A., Asadullah, A., Muddu, B., and Bhat, V. (2013). Discovery of technical expertise from open source code repositories. In *Proceedings of the 22nd International Conference on World Wide Web*, pages 97–98. ACM.
- von Ahn, L. (2006). Games with a purpose. *Computer*, 39(6):92–94.
- von Ahn, L. (2009). Human computation. In *Design Automation Conference, 2009. DAC’09. 46th ACM/IEEE*, pages 418–419. IEEE.
- von Ahn, L. (2013). Duolingo: learn a language for free while helping to translate the web. In *Proceedings of the 2013 international conference on Intelligent user interfaces*, pages 1–2. ACM.

- von Ahn, L., Blum, M., Hopper, N., and Langford, J. (2003). Captcha: Using hard ai problems for security. In Biham, E., editor, *Advances in Cryptology 'EUROCRYPT 2003*, volume 2656 of *Lecture Notes in Computer Science*, pages 646–646. Springer Berlin / Heidelberg.
- von Ahn, L. and Dabbish, L. (2004). Labeling images with a computer game. In *Proceedings of the SIGCHI conference on Human factors in computing systems, CHI '04*, pages 319–326, New York, NY, USA. ACM.
- von Davier, A. A. (2010). *Statistical models for test equating, scaling, and linking*. Springer Science & Business Media.
- Vrandečić, D. and Krötzsch, M. (2014). Wikidata: a free collaborative knowledgebase. *Communications of the ACM*, 57(10):78–85.
- W3C (2015). Semantic web - ontologies. *W3C- Semantic Web*.
- Wang, C., Chakrabarti, K., Cheng, T., and Chaudhuri, S. (2012). Targeted disambiguation of ad-hoc, homogeneous sets of named entities. In *Proceedings of the 21st international conference on World Wide Web*, pages 719–728. ACM.
- Wang, H., Zhang, K., Liu, Q., Tran, T., and Yu, Y. (2008). *Q2semantic: A lightweight keyword interface to semantic search*. Springer.
- Weber, S. (2004). *The success of open source*, volume 897. Cambridge Univ Press.
- Wei, W., Barnaghi, P. M., and Bargiela, A. (2007). The anatomy and design of a semantic search engine. *Rap. tech. East Lansing, Michigan: Department of Computer Science, School of Computer Science, University of Nottingham Malaysia Campus*.
- Wellman, B., Haase, A. Q., Witte, J., and Hampton, K. (2001). Does the internet increase, decrease, or supplement social capital? social networks, participation, and community commitment. *American behavioral scientist*, 45(3):436–455.
- Wieringa, R. J. (2014). *Design science methodology for information systems and software engineering*. Springer.
- Wilks, Y. and Stevenson, M. (1997). Sense tagging: Semantic tagging with a lexicon. *arXiv preprint cmp-lg/9705016*.
- Wolfram, S. (2009). Wolfram—alpha. *On the WWW*. URL <http://www.wolframalpha.com>.
- Wu, H., Zubair, M., and Maly, K. (2006). Harvesting social knowledge from folksonomies. In *Proceedings of the seventeenth conference on Hypertext and hypermedia*, pages 111–114. ACM.

- Xu, S., Bao, S., Fei, B., Su, Z., and Yu, Y. (2008). Exploring folksonomy for personalized search. In *Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*, pages 155–162. ACM.
- Yan, T., Kumar, V., and Ganesan, D. (2010). Crowdsearch: exploiting crowds for accurate real-time image search on mobile phones. In *Proceedings of the 8th international conference on Mobile systems, applications, and services*, pages 77–90. ACM.
- Yang, L., Qiu, M., Gottipati, S., Zhu, F., Jiang, J., Sun, H., and Chen, Z. (2013). Cqarank: jointly model topics and expertise in community question answering. In *Proceedings of the 22nd ACM international conference on Information & Knowledge Management*, pages 99–108. ACM.
- You, M.-y., Liang, L., Peng, J., and Chen, C.-y. (2009). Semantic information retrieval study based on knowledge reasoning. In *Fuzzy Information and Engineering Volume 2*, pages 271–280. Springer.
- Yuen, M.-C., King, I., and Leung, K.-S. (2011). A survey of crowdsourcing systems. In *Privacy, Security, Risk and Trust (PASSAT) and 2011 IEEE Third International Conference on Social Computing (SocialCom), 2011 IEEE Third International Conference on*, pages 766–773. IEEE.
- Yuen, M.-C., King, I., and Leung, K.-S. (2012). Task recommendation in crowdsourcing systems. In *Proceedings of the First International Workshop on Crowdsourcing and Data Mining*, pages 22–26. ACM.
- Zeng, K., Yang, J., Wang, H., Shao, B., and Wang, Z. (2013). A distributed graph engine for web scale rdf data. *Proceedings of the VLDB Endowment*, 6(4):265–276.
- Zenz, G., Zhou, X., Minack, E., Siberski, W., and Nejd, W. (2009). From keywords to semantic queries—incremental query construction on the semantic web. *Web Semantics: Science, Services and Agents on the World Wide Web*, 7(3):166–176.
- Zettsu, K. and Kiyoki, Y. (2006). Towards knowledge management based on harnessing collective intelligence on the web. In *Managing knowledge in a world of networks*, pages 350–357. Springer.
- Zhang, J., Ackerman, M. S., and Adamic, L. (2007). Expertise networks in online communities: structure and algorithms. In *Proceedings of the 16th international conference on World Wide Web*, pages 221–230. ACM.
- Zhao, Z., Feng, S., Wang, Q., Huang, J. Z., Williams, G. J., and Fan, J. (2012). Topic oriented community detection through social objects and link analysis in social networks. *Knowledge-Based Systems*, 26:164–173.
- Zheng, Z., Si, X., Li, F., Chang, E. Y., and Zhu, X. (2012). Entity disambiguation with freebase. In *Proceedings of the The 2012 IEEE/WIC/ACM International Joint*

Conferences on Web Intelligence and Intelligent Agent Technology-Volume 01, pages 82–89. IEEE Computer Society.

Zhou, Q., Wang, C., Xiong, M., Wang, H., and Yu, Y. (2007). *SPARK: adapting keyword query to semantic search*. Springer.

Zhou, T. (2011). Understanding online community user participation: a social influence perspective. *Internet Research*, 21(1):67–81.