

## University of Southampton Research Repository ePrints Soton

Copyright © and Moral Rights for this thesis are retained by the author and/or other copyright owners. A copy can be downloaded for personal non-commercial research or study, without prior permission or charge. This thesis cannot be reproduced or quoted extensively from without first obtaining permission in writing from the copyright holder/s. The content must not be changed in any way or sold commercially in any format or medium without the formal permission of the copyright holders.

When referring to this work, full bibliographic details including the author, title, awarding institution and date of the thesis must be given e.g.

AUTHOR (year of submission) "Full thesis title", University of Southampton, name of the University School or Department, PhD Thesis, pagination

**UNIVERSITY OF SOUTHAMPTON**

**FACULTY OF PHYSICS AND APPLIED SCIENCES**  
Electronics and Computer Science

**Using Social Data as Context for Making  
Recommendations (Semantics of People and Culture)**

by

**Salma Noor**

Thesis for the degree of Doctor of Philosophy  
March, 2013



# **UNIVERSITY OF SOUTHAMPTON**

## **ABSTRACT**

FACULTY OF PHYSICS AND APPLIED SCIENCES

Electronics and Computer Science

Thesis for the degree of Doctor of Philosophy

## **USING SOCIAL DATA AS CONTEXT FOR MAKING RECOMMENDATIONS**

Salma Noor

This research explores the potential of utilising social-Web data as a source of contextual information for searching and information retrieval tasks. While using a semantic and ontological approach to do so, it works towards a support system for providing adaptive and personalised recommendations for Cultural Heritage Resources.

Most knowledge systems nowadays support an impressive amount of information and in case of Web based systems the size is ever growing. Among other difficulties faced by these systems is the problem of overwhelming the user with a vast amount of unrequired data, often referred to as information overload. The problem is elevated with the ever increasing issues of time constraint and extensive use of handheld devices. Use of context is a possible way out of this situation. To provide a more robust approach to context gathering we propose the use of Social Web technologies alongside the Semantic Web. As the social Web is used the most amongst today's Web users, it can provide better understanding about a user's interests and intentions.

The proposed system gathers information about users from their social Web identities and enriches it with ontological knowledge and interlinks this mapped data with LOD resources online e.g., DBpedia. Thus, designing an interest model for the user can serve as a good source of contextual knowledge. This work bridges the gap between the user and search by analysing the virtual existence of a user and making interesting recommendations accordingly.

This work will open a way for the vast amount of structured data on Cultural Heritage to be exposed to the users of social networks, according to their tastes and likings.

# Contents

<b>Chapter 1</b>	<b>Introduction.....</b>	<b>1</b>
1.1	Research Overview .....	1
1.2	Research Problems.....	2
1.3	Thesis Hypothesis and Research Questions: .....	3
1.4	Vision.....	4
1.5	Motivation.....	5
1.6	Significance of Research.....	7
1.7	Thesis Statement and Contributions .....	8
1.8	Thesis Structure .....	12
<b>Chapter 2</b>	<b>Literature Review .....</b>	<b>15</b>
2.1	Where Boundaries meet: The Semantic and Social Web Culture .....	15
2.2	The Semantic Search Paradigm .....	17
2.3	Personalization and the Semantic Web.....	19
2.4	Social Networks .....	24
2.5	Web of Linked Data.....	32
2.6	Recommender System .....	38
2.7	Cultural Heritage Online.....	47
2.8	Summary of Related work .....	55
2.9	Conclusions.....	58
<b>Chapter 3</b>	<b>Problem Definition and Solution Design Issues .....</b>	<b>61</b>
3.1	Introduction.....	62
3.2	Shortfalls in Personalizing Recommendations .....	63
3.3	Proposed Solutions .....	66
3.4	Design Lessons learnt regarding Personalization .....	69
3.5	Social Profiling Attributes as a Base for Solution .....	70
3.6	Conclusion .....	72
<b>Chapter 4</b>	<b>Feasibility Studies and Preliminary Tests .....</b>	<b>75</b>
4.1	Combining Social, Semantic and Adaptive Aspects.....	76
4.2	Social Networks/Web 2.0 Data Feasibility.....	77
4.3	Finding the Right Vocabulary: A Universal Vocabulary or Specific vocabularies .....	91
4.4	The Cultural Heritage Domain Feasibility Analysis.....	96

4.5	Conclusions .....	101
<b>Chapter 5</b>	<b>Modeling User-Interest Semantics.....</b>	<b>105</b>
5.1	Introduction to User-Interest Semantics.....	106
5.2	Cheri System Plan .....	108
5.3	Knowledge Resources .....	135
5.4	Conclusion.....	137
<b>Chapter 6</b>	<b>Cheri System Design .....</b>	<b>139</b>
6.1	Introduction .....	139
6.2	<i>Cheri</i> system- Discover, Retrieve and Recommend .....	140
6.3	Conclusion.....	173
<b>Chapter 7</b>	<b>Evaluation and Results .....</b>	<b>177</b>
7.1	Introduction .....	177
7.2	User Evaluation of <i>Cheri</i> .....	178
7.3	Comparing <i>Cheri Experience</i> of Expert versus Non-Expert Web Use .....	200
7.4	Comparative Evaluation of Cheri Personalised Search System and V&A collection search System. ....	204
7.5	Summary .....	213
<b>Chapter 8</b>	<b>Concluding Remarks and Future Work .....</b>	<b>217</b>
8.1	Summary of Research .....	217
8.2	Research Impact .....	221
8.3	Contributions .....	223
8.4	Future Work .....	227
<b>References</b>	<b>233</b>	
<b>Appendix A</b>	<b>261</b>	
<b>Appendix B</b>	<b>263</b>	
<b>Appendix C</b>	<b>267</b>	

# List of Figures

Figure 1.5. A visual representations of contributions in this thesis. ....	9
Figure 2.3.2. The Growth of the Web.....	24
Figure 2.4.1. Timeline of the launch dates of many major SNSs and dates when community sites re-launched with SNS features. Reproduced from [80] and Updated.....	26
Figure 2.4.1. Facebook vs. Google+ Competing technologies and Services.....	29
Figure 2.5.1. Growth of Linked Open Data cloud: (a) July 2007, (b) April 2008, (c) September 2008 and (d) July 2009. (Bizer, 2009). ....	33
Figure 2.5.2. Growth of Linked Open Data cloud: (e) September 2010, (f) September 2011. ....	34
Figure 2.7.1.1. Percentage distribution of implementation choices in personalization among the studied systems.....	53
Figure 4.3.1.1. Successful DBpedia queries per category vs. % Ambiguity per category .....	95
Figure 4.4.2.1. Successful AAT and DBpedia queries per category vs. % Ambiguity per category.....	97
Figure 5.2.1. System Design.....	110
Figure 6.1. Overview and relation of the two systems .....	140
Figure 6.2.1(a). Allowing <i>Cheri</i> to extract user interests from facebook.....	142
Figure 6.2.1(b). Options for users to give access of different parts of their profile to <i>Cheri</i> application.....	142
Figure 6.2.1(c). Removing <i>Cheri</i> from user profile.....	143
Figure 6.2.2. <i>Cheri</i> Welcome Page with <i>Cheri</i> Project introduction for the user.....	143
Figure 6.2.3. V&A recommendations.....	144
Figure 6.2.4. Open-Linked-Data recommendations .....	145
Figure 6.2.5 Geo heritage.....	146
Figure 6.2.7. Facebook Open Graph.....	147
Figure 6.2.8. <i>Cheri</i> Facebook login .....	149
Figure 6.2.9. <i>Cheri</i> Website Facebook Data Extraction Request .....	149
Figure 6.2.12. <i>Cheri</i> website recommendation view .....	150
Figure 6.2.2.1. Raw interest data from facebook.....	151



Figure 6.2.2.2. wiki URI resolving/finding script using Google.....	153
Figure 6.2.4.1. Cheri Active Consumer Interest based Visualizations .....	155
Figure 6.2.4.2. Example of Linked open data recommendation result visualization...	156
Figure 6.2.4.3. Example of V&A data result visualization. ....	157
Figure 6.2.4.4 Cheri Active Consumer Map based Result Visualizations .....	157
Figure 6.2.4.5. Example of V&A Product based visualization. ....	158
Figure 6.2.4.6 Example of V&A active user location based result visualization. ....	159
Figure 6.2.5. Searching for modern pure-PHP RDF APIs (Marius Gabriel Butuc, 2009) .....	160
Figure 6.2.6 An overview of the <i>Cheri</i> Recommender System .....	164
Figure 6.2.7.1. Workflow of Cheri Search System .....	171
Figure 6.2.7.2. Example Screen-shot of Cheri Search System .....	172
Figure 7.2.1. Measure of importance of privacy in social networks .....	182
Figure 7.2.2. Measure of user experience with using handheld museum information systems. ....	182
Figure 7.2.3.1. Measure of importance of privacy in social networks. ....	188
Figure 7.2.3.2. Percentage of people have unknown friends .....	188
Figure 7.2.3.3. Type of information sharing .....	190
Figure 7.2.3.4. % Type of information sharing.....	190
Figure 7.2.7.1. Interest and Location related Geo results from V&A collection .....	198
Figure 7.3.1 (a) Trends for visiting online museums and (b) Trends for searching CH information online .....	200
Figure 7.3.2. User's experience with handheld tour guide systems .....	201
Figure 7.3.1.1 (a) and (b) Trends for adding unknown people to one's friends list.....	202
Figure 7.3.1.3 (a) and (b) Measure of personal information sharing on SNS .....	202
Figure 7.3.1.5 (a) and (b) measure of types of information shared on SNS .....	203
Figure 7.3.2.1 (a) User satisfaction distribution and (b) Usefulness measure for web recommendations (y-axis: user frequency x-axis: (a) level of satisfaction (b) level of usefulness on the scale of 1 to 5) .....	203
Figure 7.3.3.3 User satisfaction distribution .....	204
Figure 7.4.3.1. Mean precision ratios of Cheri and V&A search systems .....	211
Figure 7.4.3.2. Mean normalized recall ratios of Cheri and V&A search systems.....	212
Figure 7.5: Percent occurrence of user interests in facebook data (used in Section 7.4) .....	216

Figure 8.4.1 (b). Example CIDOC CRM Representation of the matching properties	228
Figure 8.4.1. (c). Example Mapping Process.....	229



# List of Tables

Table 2.4.1. Facebook vs. Google+ initial comparison of the competing features .....	29
Table 2.5.1. Some well-known tools currently available for creating, publishing and discovering Linked Data on the Web.....	36
Table: 2.6.1. Problems with Recommender Systems .....	42
Table 2.6.1.2. Possible and actual (or proposed) recommendation hybrids .....	46
Table 2.7.1.1 Comparative study of Implementation choices in personalization.....	51
Table 2.7.2.1. Comparative Study of Pervasive Access in Museum and Tour Guide Systems .....	54
Table 4.2.1.1. Facebook critique issues and amends a brief history.....	79
Table 4.2.1.2. Justification for using facebook for user data mining.....	85
Table 4.2.1.1. Concerns with Open Graph Protocol .....	88
Table 4.4.1.1. Search refinement using user’s social web data. ....	99
Table 5.2.1. (a) List of information elements extracted from a facebook profile.....	113
Table 5.2.1. (b) List of information elements extracted from a facebook profile.....	116
Table 5.2.3.1. Reasons for choosing FOAF.....	127
Table 5.2.3.2. Properties of the CH_INTEREST.....	131
Table: 5.2.3.3. Set of Properties used in CH_ONTO.....	133
Table 6.2.5. Comparison of Semantic capabilities in RAP and ARC.....	161
Table 6.2.7.1. Example of sorted weights for “Travel” DBpedia URI, assigned to VAM objects .....	169
Table 6.2.7.2. Time required to assign weights to all V&A objects related to a particular user interest (represented through DBpedia URIs) and Auto Predicating. ....	170
Table 7.2.1.1: Subjective measures from system interoperability templates.....	185
Table 7.2.1.2. Objective measures from system interoperability templates .....	185
Table 7.2.3.1. Subjective measures for SNS trust and privacy issues .....	187
Table 7.2.3.2: Objective measures for SNS trust and privacy issues .....	189
Table 7.2.4.1. Objective measures for accuracy as a quantitative measure for <i>Cheri</i> system. ....	192
Table 7.2.6.1. Subjective measures for system adaptation of <i>Cheri</i> through user feedback. ....	193

Table 7.2.6.2. Objective measures for system adaptation of <i>Cheri</i> through user feedback. ....	195
Table 7.2.7.1. Subjective measures for location based recommendations .....	196
Table 7.2.7.2. Objective measures for location based recommendations .....	197
Table 7.2.8. Subjective measures for user preference and usability of <i>Cheri</i> system.	199
Table 7.2.9. Possible <i>Cheri</i> system benefits as suggested by users .....	199
Table 7.4.2.1. Query List.....	208
Table 7.4.2.2. The number of relevant documents retrieved.....	209
Table 7.4.2.3. Normalized Recall Ratios for the two Search systems at 3 cut-off points .....	210

# Declaration of Authorship

I, Salma Noor

declare that this thesis and the work presented in it are my own and has been generated by me as the result of my own original research.

Using Social Data as Context for Making Recommendations (Semantics of People and Culture)

I confirm that:

1. This work was done wholly or mainly while in candidature for a research degree at this University;
2. 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;
3. Where I have consulted the published work of others, this is always clearly attributed;
4. 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;
5. I have acknowledged all main sources of help;
6. 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;
7. Either none of this work has been published before submission, or parts of this work have been published as:

Noor, S., and Martinez, K., 2011. Using Social Data as Context for Making Recommendations: Semantics of People and Culture. Multidisciplinary Research Showcase, University of Southampton, 2011.

Noor, S., and Martinez, K., 2010. Mind the Gap! Bridge the Distance. HIT 2010, University of Warwick, UK.

Noor, S., and Martinez, K., 2009. Using Social Data as Context for Making Recommendations: An Ontology based Approach. In, 9th European Semantic Web Conference, Workshop on Context, Information And Ontologies (CIAO2009), Crete.

Noor, S., and Martinez, K., 2009. Using Social Data as Context for Making Cultural Heritage Recommendations: An Ontology based Approach. At InterFace 2009, 1st National Symposium for Humanities and Technology, University of Southampton, UK, 09 - 10 Jul 2009.

Signed: Salma Noor

Date: 3<sup>rd</sup> March 2013.



# Acknowledgements

I would like to thank and acknowledge the following people:

My supervisors Dr Kirk Martinez for his outstanding support and assistance, throughout this PhD. I hope we will continue to work together and make significant ground in this area.

My Examiners, Dr Michael Oakes and Prof Paul Lewis for taking the time to understand and appraise my work.

My Advisor Dr. Nick Gibbins for his support when most needed. Dr Mark Weal for his discussion and assistance during the process of designing the user evaluations of the Cheri System.

Victoria and Albert Museum London for allowing the use of their online collection through the API.

Higher Education Commission of Pakistan and Frontier Women University Peshawar for funding my PhD.

The participants of my experiments, particularly those who took the time to comment and discuss the user issues and improvements for the Cheri system.

The Intelligence, Agents and Multimedia group, the new Web and Internet Science Group and the Department of Electronics and Computer Science at the University of Southampton for providing a research environment like no other. Particularly Dr. Melike, Dr. Khana, Dr. Laila, Dr. Areeb and Dr. Aza for their advice and council during the PhD and making this experience a memorable one for me.

My husband Dr. Mohammad Adil deserves a special thanks and recognition here for his everlasting support, love and helpful discussions along the course of my PhD.

My parents Noor and Zartashia for their love and prayers. My elder brother Badar for believing in me more than anyone else. My younger siblings; Flt. Lt. Qamar and soon to be Doctors, Rabia and Fakhar for supporting me through the challenges and achievements of this work.





*To my family.*



## Definitions and Abbreviations Used

AAT	Arts and Architecture Thesaurus
AH	Adaptive Hypermedia
API	Application Programming Interface
Atom	The name Atom applies to a pair of related standards; The <i>Atom Syndication Format</i> , an XML language used for web feeds, and the <i>Atom Publishing Protocol</i> ( <i>AtomPub</i> or <i>APP</i> ) a simple HTTP-based protocol for creating and updating web resources.
CH	Cultural Heritage
CRM	Conceptual Reference Model
CRUMPET	Creation of user friendly mobile services personalised for tourism
DBpedia	A data set that provides metadata about Wikipedia recourses.
DPWG	Data Portability Work Group
FOAF	Friend of a friend ontology helps describe people
GATE	An open source text engineering architecture for extracting named entities from documents
GUIDE	Is an intelligent electronic tourist guide developed to provide tourist information about the city of Lancaster.

HCI	Human Computer Interaction
HIPS/Hippie	(Hyper-Interaction within physical space) Is a nomadic system aimed at allowing people to navigate through the physical and the related information space simultaneously, with a minimum gap between the two
IE	Information Extraction
IEEE PAPI	IEEE Public And Private Information for Learners – is a data interchange specification developed for communicating between different systems
IMS LIP	IMS Learner Information Package Specification – is a specification recording lifelong achievements of learners and transfer of these records between institutions
IR	Information Retrieval
JSON	JavaScript Object Notation is a lightweight data-interchange format. It is easy for humans to read and write. It is easy for machines to parse and generate. It is based on a subset of the JavaScript Programming Language, Standard ECMA-262 3rd Edition - December 1999. JSON is a text format that is completely language independent.
LD	Linked Data
LOD	Linked Open Data
KB	Knowledge Base
MOAT	Meaning of a tag ontology as the name indicates is a collaborative framework to help Web 2.0 users give meanings to their tags in a machine readable format.
OWL	Web Ontology Language
SN	Social Network
SNS	Social Networking Site

RDF	Resource Description Framework
RDFS	Resource Description Framework Schema
RESTful	Representational State Transfer (REST) is a style of software architecture for distributed systems such as the World Wide Web. REST has emerged as a predominant Web service design model. A RESTful web service (also called a RESTful web API) is a web service implemented using HTTP and the principles of REST.
SCOT	Social semantic clouds of tag ontology presents a model for main properties and concepts required to represent a tagging activity on the semantic Web.
SICO	Ontology to define main concepts that are needed to describe a social community on the semantic Web. The aim is to view a person's entire contributions on the social Web.
SKOS	Simple Knowledge Organization System
SW	Semantic Web
UIH	User's interest hierarchy
URI	Uniform Resource Identifier
V&A	Victoria and Albert Museum London
W3C	The World Wide Web Consortium
WWW	World Wide Web
XML	eXtensible Markup Language



# Chapter 1

## Introduction

### 1.1 Research Overview

The World Wide Web is a global information space on an unprecedented scale. Recent years have seen exponential growth of social network sites (SNSs) such as Facebook, MySpace, twitter, LinkedIn and Google+ to name a few, which have attracted hundreds of millions of Internet users over the last few years. This research hypothesises that, by utilising the vast amount of user generated information enclosed in online social networks and by making it reusable, a richer and more dynamic model for managing user interests can be achieved. Such an *interest profile* has varied applications in domains including recommender systems, online search systems, personalised information retrieval and in general any services that deal with context sensitive and user adaptive processes. Our work is an initial effort in modelling such a mechanism for recommending and searching personalised cultural heritage information online.

The *Cheri* system is a user interest capturing, profile generating and art recommending system designed to make the Cultural Heritage domain more reachable to the general Web user. The Cheri system makes use of the vast amount of user generated data on the SNS Websites to identify their interest. The interest profile generated by *Cheri* is mapped through LOD standards which make it reusable across the Web as well as machine readable. The interest profile is however layered with a mapping layer to provide multi-domain knowledge. The system uses the interest profile to recommend artwork from the art collection of the Victoria and Albert Museum, London that currently contains over a million records (V&A Search the collection, 2011), as well as open source information from DBpedia and the Web. The



mapping layer helps facilitate this process by interlink the user interest terms from the user profile with the appropriate concepts in the museum records using a multi-domain ontology.

## 1.2 Research Problems

The main research issues that the system will solve include:

- Avoiding the cold-start problem.

*The Cold start problem is a common problem in personalised recommender systems and its root cause is lack of user interest information and or ways of capturing it. The problem of finding and updating user interest information unobtrusively and dynamically while relating it to appropriate concepts to suggest relevant information resources is still not solved.*(this is discussed in greater detail in chapter 3)

- Sub-problems of cold-start problem:

- Shifts and temporal cycles of user interest.
- Recommendations made independently of context.
- Only items identified in one pre-specified representation are considered.
- The most similar items are not always good recommendations.

All of these are well known problems in personalised search and recommender system research. Details and a discussion on these can be found in chapter 3.

- User Interest Capturing: Efficient, unobtrusive, self-sustaining and dynamic approach to user interest capturing.
- Data Filtering and Concept location: A comprehensive vocabulary for describing and annotating social network data.
- User Interest-Profile modeling: A portable, reusable and machine readable way of representing user interest information.

- Using the interest Profile in *Query optimization* and *Result filtering*.
- Testing the effectiveness of recommender system in terms of its adaptation quality.
- Testing the quality of recommended items (accuracy, novelty, enjoyability) relative to the users experience (time taken to register and time to recommend).

### 1.3 Thesis Hypothesis and Research Questions:

To address these issues in this research we propose the following set of hypothesis;

#### **Hypothesis#1**

*‘When the user is not asked to enter too much information about themselves and their interests to boot-start the recommendation process in a system, rather the system acquires it through users social networking activities, this can decrease the effort spent by the user, increase the ease of use of the system and help solve the cold start problem’*

#### **Hypothesis#2**

*‘Social web data can be used to gather up-to-date interest information about a user. The user’s SNS interaction activities will better represent the user’s ever changing interests.’*

#### **Hypothesis#3**

*‘Ambiguity of SNS data can be clarified if their context is well defined and standard vocabularies and ontologies are applied to resolve this issue.’*

#### **Hypothesis#4**

*‘A generalised user interest-profiling system Based on users SNS data can serve as an interpretation of user’s interest and assist during recommendation or searching processes.’*

## **Hypothesis#5**

*‘The profile thus generated will represent interests as concepts in a standard ontology and can serve as a useful resource for the recommender system in determining user’s interests and possible intentions while making recommendations, and in designing a mechanism for automated query formulation through the use of SN data.’*

## **Research Questions:**

Based on the above mentioned research problems and hypothesis this thesis asks the following questions and provides answers to them in the design, implementation and evaluation of the Cheri system.

- Whether it would be easy to capture the user interest data from a SNS and if so will the user find the process easy?
- Will the interest transfer from the users SNS be annotated and identified with the right concepts semantically?
- Whether LOD is going to be an effective way to bring our cultural heritage resources like libraries, archives and museums together in the open as a fully connected and integrated source of knowledge?
- Can SN help users find interesting contents on CH Websites?
- How can user interest information obtained from SN lead to better recommender design?

All of these are well known problems in personalised search and recommender system research. Details and a discussion on these can be found in chapter 3.

## 1.4 Vision

The vision of this research is a novel idea of “a walking museum rather than a walk-in museum”. We envision a museum, which brings Art to it’s admirers and Artwork to its viewers rather than the other way around. A walking pervasive museum is a museum that brings art to the visitor wherever they are but in a highly personalised manner. A museum that describes history to its visitors where history was made, a

museum that treads in the same footsteps of the emperors and paupers alike to tell the most wonderful yet fact based story, that is, our shared history.

While the world is making a history of its own the virtual world which is a very real world in our daily lives is now taking a leap from just being a data repository and a communication medium to being a knowledge resource that can be made intelligent enough to help and assist its users in exploring and rediscovering the resources that it carries. Big yet very possible visions like the semantic Web, the Web of LOD and now the Web science initiative (Hendler, et al., 2008) are exploring the capabilities of this not-just-a-data-resource (called Web) in a very real way or rather a very natural to the machine way, by enabling a means of representing what is human-readable as machine-readable. While these advancements are here and possible, it would be a shame to have this ever-growing resource called the Web and not use it to enrich and facilitate our traditional yet very profound resources like the museums.

The Web research today can make history in making the history of the world almost like a virtual/pervasive experience for its users. Virtual experience can be defined as experiencing and visiting a place that you have not been to in real life. But *virtual* here is being in the same place where the history has occurred and experiencing it through our great knowledge resources like the Web and the museums in a very personal manner- *almost time travel*.

## 1.5 Motivation

It is said dreaming is a bad habit, though the world of Science and the world of Arts both stand on the shoulders of great dreams realised. I believe dreaming is a bad habit if you don't have the courage to believe in them and the hope that they might come true. My motivations for this research are two such inspirational dreams dreamt by two great people and realised by the world around them.

The first dream is of a drummer in the first band (Severe Tire Damage) to perform live on the internet. Born in the suburbs of Chicago, he dreamt of a world where the technologies weave themselves into the fabric of everyday life until they are indistinguishable from it. This man whom we now know as the father of Pervasive Computing, Mark Weiser told the world to "say goodbye to your computer -*it's about*

*to disappear*. That is, it will be so much a part of your life that you won't even know it's there" (Weiser, 1991).

The Second dream is of an English boy who in his university days was caught hacking and was subsequently banned from using university computers. In later years his achievements as a computer scientist marked him as 1st in the 100 greatest living geniuses (The Telegraph, 2007). There are many things that Sir. Tim Berners-Lee has done to contribute to the world but the most important of his inventions is the World Wide Web. Now his dream is to make his invention of the Web to reflect "*a vision of the world. A truly connected world*", which he named the Semantic Web (Berners-Lee, 1998). This Semantic Web is not a separate Web but an extension of the current one, in which information is given well-defined meaning, better enabling computers and people to work in cooperation.

Looking close enough, both the visions mould into the same reality as indicated in his views on the Mobile Web initiative by Sir. Tim Berners-Lee that, *information must be made seamlessly available on any device* (Berners-Lee, 2006). If the semantic Web is a reflection of the world it should be accessible from anywhere any time and if the technology really is to weave itself in the fabric of everyday life it needs to understand the world of humans as humans do, in a very personalised way, which can be made possible by the semantic Web technologies.

Our research interest lies where the two dreams meet. A third important ingredient in our equation of inspiration is the fabulous social Web revolution that we observed during the second half of the last decade. The semantic-Web and social-Web are two very different entities and where they meet in harmony they give life to the Web 3.0 revolution (Shannon, 2006) which aims to link the knowledge and expressiveness of the two domains. Their unification is an interesting arena full of possibilities on the individual as well as the community level.

Our work is an effort in the same direction. We propose a system that gathers information about the user's interests from the social Web and enriches it through semantic Web technologies. Hence it creates an interest model profile for the user with the help of the best in both the technologies, which can serve as a rich context base for search and retrieval systems. It finally queries over an open corpus (linked-

data on the Web) as well as a considerably closed corpus semantic data source (museum repository) to make its recommendations.

This work investigates the use of social Website information to enrich the contextual information used by systems with user interests making recommendations, especially within semantically enabled knowledge.

## 1.6 Significance of Research

Our research is stirred by the following motives, listed under their respective research themes:

### 1- Data Filtering/ Data Mining

Increasing requirement for global access to the wealth of highly distributed, heterogeneous and dynamic content has led to the need for integrating information from multiple resources. This means more information and an accentuation of the data overload problem. However, part of the problem is being answered by the use of research in information retrieval and data mining, through provision of mechanisms for text analysis and standard data filtering. Information seekers will understandably appreciate a filtering mechanism that could further exclude resources that are irrelevant to them while at the same time make available high quality and useful items most relevant to the user's requirement. In our research we will adopt the standard vocabularies and an ontology based approach for tailoring to an individual's user requirements.

### 2- Understanding the User/ User Modeling

One of the challenges in understanding the user better, is to uncover the latent semantic boundaries that frame human and cultural communities; it is interesting to note here that the semantics of a single term or concept may be perceived differently in different cultures depending upon the use of that particular term in the context of that culture. Such differences may not be easily visible and thus require special attention and in-depth knowledge so that we can design search systems to better suit people's natural expressive tendencies. We also wish to deliberately designing systems to leverage the inherent semantic mechanism that guides how people

naturally want to conceptualise, interact and communicate. Existing user modelling standards, IEEE PAPI (IEEE, 2000) and IMS LIP (IMS, 2005), are mainly developed to elicit a user's learning requirements and are thus insufficient for Web based personalization. New user models are needed to support the user's recreational and aesthetic needs (in general the user's interests). Thus we introduce a user interest based profile model for Web search personalization.

### 3- Social and Semantic Web

It is evident from the trend through the last decade that even when mountains of technological standards were complete, Web 2.0 and social media did not really take off until the activities of the semantic production- such as tagging, rating and associating- became easy, transparent and rewarding enough to sustain organic growth of participation. The same can be said for the semantic technologies and their mainstream adaptation. Our research utilises the best of both worlds to resolve the issues in user adaptation and personalization across search and recommender systems.

### 4- Linked Data

With time and effort the linked pool of open source metadata over the Web is growing and becoming a prominent, if not significant, entity. The possibilities to exploit this information are many. Our research investigates the possibilities of utilising and incorporating LD on the Web and how the users can benefit from it without the full-fledged adoption of the semantic Web vision. LD provides new possibilities for open-corpus search, by dynamically relating user models to any dereferenceable URI.

## 1.7 Thesis Statement and Contributions

A framework will be developed to enable Cultural Heritage related Personalised Recommender Systems to consider Social Networking Data for dynamic user interest profile generation. Such a framework will contribute towards reducing the semantic gap between the cultural heritage expert domain knowledge and general Web user interests. Moreover, as a consequence of the cross domain nature of the user profile, such a system will provide recommendations that are high quality, unexpected and geared solely towards satisfying user needs.

Figure 1.5 outlines the various contributions made in support of this thesis:

#### Integrating Domain Resources, expert knowledge and user profiles

- Dbpedia used as a universal vocabulary (expert knowledge) for defining concepts in user SNS data.
- Each resource is mapped to a corresponding concept in DBpedia Ontology
- Relationship between the resource and Dbpedia concept is identified by URI of the Wikipedia entry describing that concept
- Relationship between the user interest and Dbpedia concept is identified by URI

#### User profiling

- Unobtrusively collect information about user interests from their social network profile
- Generate a FOAF based user profile annotated with DBpedia ontology concepts
- Portable, Reusable, Dynamic and automatically updated user profile model

#### Recommendation Context

- context provided as current state of the user when the recommendation was needed e.g., current location
- Defined as a subset of user profile, relevant to the recommendation need.
- Alternatively obtained by specifying the search term for which the recommendations are sought.

#### Novel filtering framework for recommendations

- Dynamic generation of user's current top 5 resource types in the current domain, calculated through user's current interests.
- Dynamically updated and recalculated for every new search session

#### Evaluation

- Evaluating the quality of recommendations from a user's perspective

Figure 1.5: A visual representations of contributions in this thesis.

The contributions of this thesis are listed as follows:

- 1- Our research has helped build a personalised cultural heritage search and recommendation system using strong semantics supported by standard semantic and social Web technologies, utilising the social Web as a context source. This generalised user interest profiling model helps the research system (*Cheri*) keep track of the changes in the user's interests over time and incorporates these changes in the current search context accordingly and hence aid personalised recommendations while avoiding some of the most well-known pitfalls in recommender systems as discussed in chapter 3.



- 2- Building an exportable interest-profile for SNS users which is dynamic and designed using the W3C standards which makes it reusable and easily extendable if required.
- 3- Our portable interest model will contribute towards a unified user experience across different sites, easy information access for service providing agents like recommender systems and end-user applications, increased recommender productivity due to less time required to search user related information (such as user interests .), better planning of retrieval strategies and more accurate evaluation, better equipped exchange of user information across different platforms and above all meaningful personalization.
- 4- ***Introduce a fresh approach to solving the well-known cold start problem.*** The Cold start problem (Lam, et al. 2003) is rendered as the main problem to be solved in the context of the proposed framework. This is achieved by ensuring that users are not assigned empty profiles upon registration, but rather carry with them the information that reflects their current interests across multiple domains. Of course, if users have not created any information prior to subscribing with the system (or have chosen to not disclose any) the problem persists. However, such behaviour would somewhat defeat the point of seeking personalised recommendations. The user interest information is automatically gathered from the user's social networking account and is dynamically updated. To avoid initial and constant updating efforts required in making the user profile, linking the user interest model with the user's SNs profile is implemented as a solution.
- 5- ***Improve findability and resolve the item similarity issue in recommender systems:*** The Cheri framework requires each resource to be mapped to a unique set of terms in the universal vocabulary (DBpedia). This provides a **mechanism for identifying interchangeable resources**. Such resources are expected to have identical descriptions using terms from the universal vocabulary (DBpedia) and can therefore be merged. This mechanism calculates the equivalence amongst items which is the basic solution for item similarity issues; and increases the 'Findability' of previously hidden yet related information aiding in **new knowledge discovery**.

6- ***Our novel interest filtering technique*** implemented in the Cheri system (besides solving the problem of shifts and temporal cycles of user interests) also has shown promising results in helping to elevate the ‘similarity of item’ problem common in recommender systems, by ensuring that the recommended results are always from a set of highly weighted resources across a set of resource types (best representing the user’s current interest) rather than from a single type of resource. In our proposed filtering model (see Chapter 6, Section 6.2.6) the query results are presented in order of relevance, but to avoid the “*most similar items are not always good recommendation*” phenomenon (explained in Section 3.3.2), our novel approach of filtering the results thus obtained with the current top five resource types (from the domain) that the system has calculated to be most related to the user current interest profile, has shown promising results, the approach is further explained in detail in section 6.2.6. This has succeeded to bringing variety in the results without losing relevance to the user, as can be seen from the results of the evaluation in section 7.4.3. In addition the automatic upgrading of interest profiles each time a user logs new interest in their SN ensures the dynamic interest profile that forms the core for the recommender system keeping sure that the current interest is always considered while handling a user search or query string.

7- The Cheri framework successfully presents a ***novel solution for potential biasing effects in recommender systems*** by shifting the emphasis to satisfying user needs. By introducing a standalone user preference/interest calculation and updating mechanism independent of the end data resources, it becomes harder to spuriously insert an arbitrary recommendation. Moreover, to influence the system to recommend a biased resource over others, one would also have to obtain control over the universal representations of resources and the semantic connections between their descriptive terms in the universal vocabulary. Furthermore, since the SN data are simply seen as platforms indicating the preferences of their members, there is no guarantee on what objects will be selected for a user, on the bases of extracted SN data, as a recommended resource.

- 8- On its launch, neither Facebook nor the publishers (its partners) did any mark-up on their pages. At the time none of the entity pages on Facebook.com had Open Graph mark-up and thus Facebook's own pages remain closed. Ironically, this might not be because the company does not want to mark-up the pages, but it might be because it can't until it figures out what is actually on the page. This is what semantic technologies have been working on over the past several years. In this thesis we ***introduce a feasible way of marking up user data on the Facebook graph via a universal vocabulary (DBpedia)***, though not unique to semantic search research, it would be the first time to suggest it as a solution for a big SN graph.
- 9- In the issues with recommendations made independently of context it is realised that the object/resource attributes alone are not adequate for representing the context of a recommendation. The framework offers a solution for automatically determining which aspects of a user interest profile are relevant to the context of a particular query. It achieves this by providing a ***novel search tool*** which overlays the user's current interest rating with the context of the user query to produce results explicitly selected to reflect a particular context. The search results are filtered through a user interest matrix. Any resource type that the framework finds are related (through semantic annotation and ranking) to a user interest, that is, the user has implicitly expressed interest in it as part of their profile, regardless of their origin, is considered. By adopting this mechanism, as such, the effects of problems associated with the inadequacy of user profiles to represent a wide range of user interests are expected to be less severe.
- 10- Helping to make the CH resources online more approachable for the users of SNSs.

## 1.8 Thesis Structure

The rest of the thesis is structured as follows:

Chapter 2: *Literature Review*: presents the background of this research. The chapter gives a history of Semantic search and Personalization. It describes what part the Social and the Semantic Web can play in conjunction with each other. A history of personalization techniques developed in the Adaptive Hypermedia and the Data

Mining communities are discussed farther on and an in-depth look at the techniques developed in these communities is provided. A literature review of Social Networks and Web 2.0 is presented next discussing the history of social network sites and their part in the evolution into Web 2.0 and Web 3.0, as well as previous research about social science and recent research on SNSs. This provides a historic framework and evidential materials for our research on online social networks as a substantial resource for user interest profiling. Next we present a literature review of the Recommender system domain and place our research in the light of the state of the art in this field. Next is an in depth analysis of the museum and tourist domain systems and the personalization they have provided over the years. This chapter also presents a case study on the CHIPS museum project and concludes with an in-depth survey of the personalization and pervasiveness facilities provided in both commercial as well as research developed tourist and museum recommender systems. A related work section gives an up-to date discussion of the systems with similar research ventures. Finally we state our research hypotheses based on the research questions raised in chapter 1.

Chapter 3: *Problem definition and Solution design Issues*: discusses and refines the problems identified by this research and discusses possible directions to solve these problems, suggesting that social network data can form the bases of these solutions.

Chapter 4: *Feasibility Studies and Preliminary Tests*: describes the initial work we undertook to justify the use of Social Networks/ Web 2.0 data as a source of gathering user interest information, more precisely a discussion on Facebook as a Social Network of choice for this type of user data mining. The study to identify the most appropriate vocabulary for the type of data we were extracting is discussed next. The justification and reasoning behind using a universal vocabulary rather than a more specific vocabulary is discussed. Finally finding and testing the feasibility of the structured or semi-structured cultural heritage data online that we used to generate the knowledge base for artwork recommendations is presented.

Chapter 5: *Modeling User Interest Semantics*: describes our work in semantic modelling of user interest information extracted from Web 2.0 sites. The requirements identified for adding semantics to user interest representations are presented. Next an introduction to the proposed framework design for a personalised recommender

system that fulfils the requirements is presented. Integrating the user interest model with the recommender system and annotating user interest information from Web 2.0 sites with LD information resources for providing personalised recommendations are discussed.

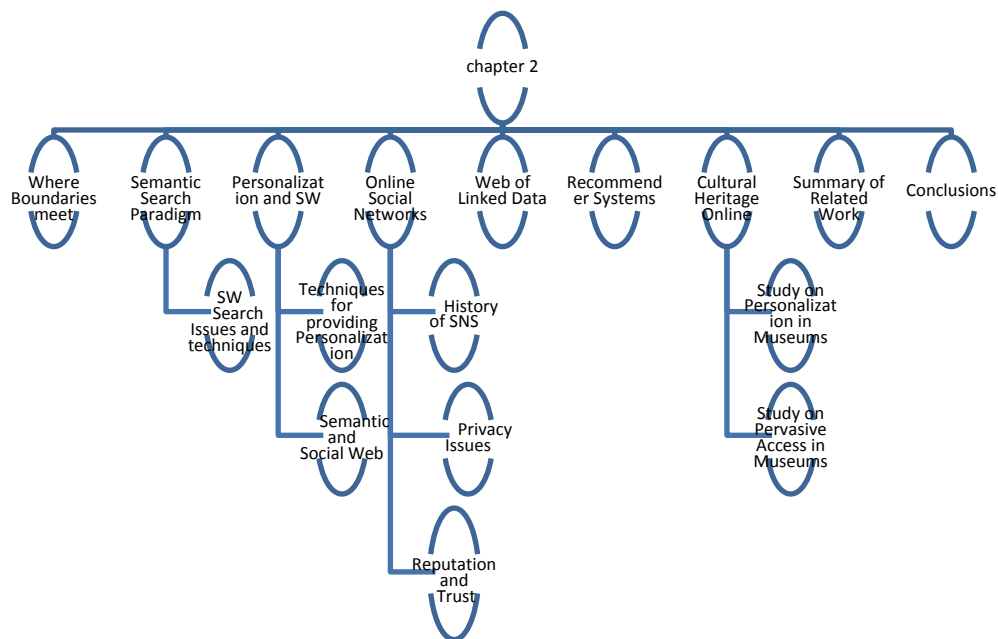
Chapter 6: *Cheri System Design*: explains our personalised search and recommender systems. The chapter describes the general architecture of the systems and the technologies used. Their functionalities are explained and discussed in detail. The technical details of the implemented parts are given by example.

Chapter 7: *Evaluation and Results*: present the design for the evaluation of the two *Cheri* systems. We test the infrastructure and examine the usability of the two systems. A user evaluation of the Cheri recommender system is first discussed. This evaluation is conducted as a proof of concept for our thesis statement presented in Chapter 1 section 4.1. It further on tests the Cheri system across 8 different parameters which justify our *vision* and the *thesis contributions* as discussed in Chapter 1. Furthermore this chapter also presents the results obtained from two other evaluations. One is a comparative user evaluation between expert and normal Web users and gives useful insights in the usability of the system and the second is a comparative evaluation of the Cheri search system and the Victoria & Albert search the collection online system. For each evaluation, the methods and experimental results are described.

Chapter 8: *Concluding Remarks and Future Work*: finally this chapter summarises our work and presents future perspectives for this research.

## Chapter 2

# Literature Review



Chapter 2 topic hierarchy

*Believing that a connected Web is a smarter Web, the Semantic Web goal goes far beyond building a Web of machine-processable data, aiming at enriching lives through access to information (Davis, 2009).*

### 2.1 Where Boundaries meet: The Semantic and Social Web Culture

This chapter discusses the literature review of the research areas and technology that we base our research on. We start by discussing the search paradigm as we look forward to the next stage in the evolution of the Web- the semantic Web. The aim of the semantic Web is to shift the emphasis of the association and linking from documents to data. This will pave the way for a more comprehensive form of reasoning. The change will have three main positive impacts as stated by (Hall, 2011).

It will assist data re-use, often in an unexpected and new context. It will reduce the cost of relatively expensive human information processing and finally it will open up vast amounts of information trapped in relational databases and Excel spread-sheets by making it machine readable.

Next is a discussion on the Social Networks and other Web 2.0 technologies and how the union of the semantic and social Web has paved the way for Web vision 3.0. The success of the Web 2.0 technologies like social networks is characterised by the simple approach they adopt for sharing information on the Web. Over the years the use of the Web has drastically changed and this change facilitated by the semantic and social Web technologies has brought revolutionary changes. The initial use of the Web can be characterised by most of the Web users consuming content that a comparatively small set of developers had created. With the introduction of social Web technologies everyone has become, as George Ritzer and his colleague claimed a *prosumer* (Ritzer and Jurgenson, 2010), that is someone who produces and consumes content. The reason for the success of such tools and services for Web cooperation and resource sharing is characterised by their ease of use, but as these systems grow large the users feel a need for more structure to better organise their resources and to help enhance search and retrieval. The answer to these concerns we believe lies with LD. Next is a discussion on the Linked Open Data initiative and its relevant concepts, the progress it has seen over the years and the different tools and services that have emerged in support of this LD initiative.

The Semantic Web encapsulates a vision of a Web of LD. It helps automate or semi-automate querying of data from heterogeneous resources on the Web and facilitates sharing and interpreting it (Shadbolt, et al., 2006). Thus the needs of the Web 2.0 services and tools very much benefit the services provided by the Web of LD. The basic building blocks here are the Universal Resource Identifiers (URI), Resource Description Framework (RDF) and Ontologies. It is hypothesised (Hall, 2011) that LD can become the domain data sharing and data integration platform and its effect on society and the way we use the Web will be profound.

Next we give an overview of the Recommender System research and where our research can be placed in the area. The main research problems that are shared across this domain are discussed, and an overview of the state of the art in the research done in solving the information overload, personalization and other problems is given.

Finally cultural heritage online, being our target domain for this research is discussed. A study on the state of the art in Personalization in museums and tourist domains across the literature is carried out and a similar study for the provision of pervasive access in cultural heritage is discussed next.

The chapter concludes with the state of the art in related areas.

## 2.2 The Semantic Search Paradigm

The Semantic Web (SW) is the vision of the next generation of the Web as proposed by Tim Berners-Lee (1998), where the meanings of the Web contents hold more importance than ever before. Semantic Search seeks to improve the accuracy of a search by understanding the intent of the user and the contextual meaning of the terms as they appear in the searchable data-space. The search space can be a closed system or the World Wide Web. However, Makela (2005) defines Semantic Search as search requiring semantic techniques or search of formally annotated semantic content.

### **Semantic Web Search Issues and Techniques**

The semantic search employs a set of techniques like using ontologies for retrieving knowledge. Some basic research issues in the semantic search domain as listed in the literature include;

- Augmenting traditional key word search with semantic techniques
- Basic concept location
- Complex constraints queries
- Problem solving and
- Connecting path discovery

Most of the research done on the issue of *Augmenting traditional key word search with semantic techniques* did not assume the knowledge to be formally annotated.



Instead the ontology techniques were used to augment key word search either to increase recall or precision. Most of these implementations used thesaurus ontology navigation for query expansion. Examples include (Buscaldi, 2005) and (Moldovan, et al., 2000) who used the approach of expanding terms to their synonyms and meronym sets using the Boolean OR operation. In Clever Search (Kruse, et al., 2005) the user selected word senses of the corresponding term in the WordNet ontology is Boolean ANDed to the search term to clarify the semantics of the query. In (Guha, et al., 2003) terms are matched against concepts in an RDF repository besides being used in the normal keyword text search. The matching concepts from the RDF repository are returned alongside the document search, the idea here being not to expand the search terms, but rather to annotate the documents with the related concepts. Rocha, et al. (2004) uses an RDF graph traversal method to find related information in the results of the keyword based search query. The idea here is to annotate the document to find its relevance to the concepts. Airio, et al. (2004) uses a direct ontology based browsing interface to find relevant documents where concepts in the ontology can be selected to constrain the search.

The research done on *basic concept location* took advantage of the fact that the data that the user is searching for is usually an instance of a class in an ontology. So by sorting out the concept, individual and relationships which are the core semantic Web data types, the instances can easily be identified. So, user is presented with a general hierarchy of classes in the ontology from which he chooses a class to which the instance he is looking for belongs. Then the related properties and relationships of the class are sorted to get the desired results, for example in (Heflin, et al., 2000) and (Maedche, et al., 2001). An interesting approach is that of the “Haystack Information Management System” (Karger, et al., 2005; Quan, et al., 2003), who designed their user interface almost completely on browsing from resource to resource to locate the desired concept. A similar idea is discussed in the research of Teevan, et al. (2004) who argues that mostly the user has an idea of the related resources to what they are actually looking for rather than knowing the specific qualities of what they are looking for. Thus searching in this case becomes more of a browsing experience. This idea of the related measure allows the user to partially specify their information need up front depending upon their knowledge of the target resource and enabled them to take

advantage of the contextual information they knew about their information target. The third issue is of *complex constraint queries* that are not very difficult to formalise in the semantic Web as they can be visualised as graph patterns with constraint nodes and links. But such queries are very difficult to formulate correctly from a user's point of view. Much of the research to dealing with this issue has been done in the user interface domain, for helping the user to formulate complex queries. Athanasis (2004) introduces a graphical user interface (GUI) that allows the navigation of an ontology to create graph pattern queries. Catarci, et al. (2004) present a similar approach differing in the way that the user is provided with example graph queries which they can customise according to their needs.

The fourth issue is of *Problem solving* that is describing a problem and searching for a solution by inferences based upon the ontological knowledge available on the issue. However research on this issue is very limited. Fikes, et al. (2003), describes a query language for the semantic Web based on DL-reasoner which allows simple *if-then* reasoning. A project based on this system is (Hsu, 2003).

The last issue *connecting path discovery*, focuses on the fact that a lot of interesting and useful information is encapsulated in the links between the resources rather than the resources themselves. Study is required to determine a means for discovering, inferring and extracting the information in the links. Kochut and Janik (2007) introduce SPARQLeR, which aims to add the support for semantic path query.

The methods so far used to solve these issues include RDF path traversal, mapping between keywords and formal concepts, graph patterns, logics, combining uncertainty with logic and view based search.

## 2.3 Personalization and the Semantic Web

Though the SW concept revolutionises the vision of the Web, making it more productive, it came with a package of research questions and issues, most of which the research over the last decade has tried to answer. But there has been a considerable lack of research in certain areas, personalization being one of them.

We see that the most successful semantic applications developed so far have been those designed for closed communities like employees of large corporations or scientists in a particular area, whereas applications designed for the general public are mostly prototypes or in-laboratory experiments.

Hence, for the general Web user a particular promise of the semantic Web for Personalization, Large-scale semantic search (on the scale of World Wide Web), is still in many ways unresolved. Below we will try to give an overview of the research in personalization across several research domains.

The aim of personalization is to support the user in accessing, retrieving and storing information. The provision of this support may require consideration of the user's interests, current task, the context in which the user is requesting the information, the device he is using, user disability if any, time constraint and communication channel constraint etc. As personalization requires a lot of things to consider, naturally it becomes an interdisciplinary problem and is studied in different disciplines such as, hypertext (adaptive hypermedia systems), collaborative filtering (recommender systems), human computer interaction (adaptive interfaces) and artificial intelligence.

Personalization usually requires a software system/machine to assist a human to acquire his desired results. This requires the knowledge to be interpreted in machine readable format as characterised by the semantic Web.

Personalization occurs at the ontology layer but mostly at the logic and proof layers (Baldoni, et al., 2005). If we look closely at the layered model of the Semantic Web [Figure 2.3.] we find that the semantic Web envisions an inference mechanism above the knowledge layer to act as a means of providing content-aware navigation and producing an overall behaviour that is closer to what the user desires. That is why the semantic Web is the most appropriate environment for realizing personalization.

In other words the semantic Web is deeply connected to the idea of personalization in its very nature (Baldoni, et al., 2005). This fact can be realised by studying the outcomes in a semantic Web system, where it is observed that the results are always adapted or personalised to meet specific requirements.

### 2.3.1 Techniques for providing Personalization

If we define the process of personalization as a “process of filtering the access to Web content according to the individual needs and requirements of each particular user” as defined by Baldoni, et al. (2005), then we can say that personalization is achieved through the application of various filters. The research on these filters has been done in two research areas, Adaptive hypermedia systems and Recommender systems (using Web mining techniques).

#### **Techniques in Adaptive Hypermedia Systems**

Adaptive hypermedia systems as defined by Brusilovsky (1996) are all hypertext and hypermedia systems which reflect some features of the user in the user model and apply this model to adapt various visible aspects of the system to the user. Thus, adaptive hypermedia systems provide personalization to individual users, resolving the *lost in hyperspace* problem.

Hypermedia systems consist of documents linked together in a meaningful way. Therefore two kinds of adaptations are possible in these systems to achieve personalization i.e. adaptation of document contents and adaptation of links. Document level adaptation is achieved through enriching the document with metadata and some parts of the documents even require re-writing to achieve different adaptation results.

For *document level* adaptation the techniques used to adapt the contents of a document include: Stretch text, Page or page fragment variants, Frame based techniques, Conditional text, Additional explanations, Comparative explanations, Sorting, Explanation variants and Prerequisite explanations.

However, *link level* adaptations are generally used to help personalise the navigation of the user in an adaptive hypermedia system. The techniques used to do so include direct or sequential guidance, prerequisite knowledge sorting, similarity sorting, Adaptive hiding, Link annotation and Map annotations (Specht, 1998).

No matter how the adaptation is provided to personalise the system, the techniques used so far are applicable to a certain system with a number of specifications. That is, the functionality of the system is only specified according to its particular

environment. This might be true to some extent, due to the well-known open corpus problem as described by Brusilovsky (2001) and Henze, et al. (2000) which results in a lack of re-usability or interoperability among the adaptive hypermedia systems and techniques. However, the standardization of metadata formats through the SW is a step forward in solving this problem.

### **Techniques in Web Mining**

The personalization achieved through Web mining techniques does not have at its base a well-defined corpus like in the case of the hypertext for adaptive hypermedia systems. Instead it depends upon the graph-like view of the world wide Web. As we know that the graphical view of the Web is constantly changing so a completely known structure of the Web at any time is not possible.

Thus the personalization provided by Web mining relies on the physical (hyperlinks) and the virtual (related in concept but not hyperlinked to each other) links existing among the documents (Baldoni, et al., 2005). So two approaches are used to detect the relationship among documents, i.e. mining based on *content* and mining based *usage*.

The techniques used for Mining-based personalization include Content-based recommendations, collaborative recommendations or social information filtering, demographic recommendations, utility-based recommendations and knowledge-based recommendations, as surveyed by Burke (2002). However for these techniques to work properly a considerable amount of data is needed.

As there is a lack of a well-defined structure corpus in Web mining systems, the user-modeling is usually restricted to an interest or a content profile. However, it is important to note here that user-modeling is the core for each personalization process, as the system uses it to identify the user's needs. If the user model is not correct even the best personalization algorithm will yield the desired results.

#### **2.3.2 Semantic and Social Web**

The semantic and social Web, are two very different entities. However bringing them together promises to link the knowledge and expressiveness of the two domains. Their

unification is an interesting arena full of possibilities for the individual as well as at the community level. In recent years, the introduction of APIs by several social Websites opened a way for developers to reuse vast amounts of information on the sites to experiment and produce worthwhile applications. This was also welcomed by semantic Web researchers and data from the social Websites soon became a rich test-bed for future semantic and social Web technologies. Similarly Microformats and structured blogging efforts paved the way for blogging data to be brought into the semantic Web. Amongst other useful things, one of the most interesting outcomes of this semantic and social Web merger is the possibility of utilising this huge amount of user-created data to understand the user better. Studies (Li, et al., 2008), (Golder, et al., 2006) indicate that the tagging activities of an individual carry interesting information about his/her interests and therefore can play a vital role. We believe that by linking all the different social identities of an individual over the Web and by unleashing the vast amount of tagging information enclosed in them, a richer and more dynamic model of user interests can be achieved. That can serve as a rich context to further assist adaptive and user oriented applications and search processes. Unified profiling and tag data portability efforts are a way forward in this direction. In the last few years several Web2.0 sites started to provide links to export data from other social networks and within days the social existence of a Web user became more unified, e.g. Youtube for Myspace, Digg, orkut, live spaces, bebo, hi5, mix and Facebook; Orkut for Youtube and Facebook for flickr. Similarly major internet players like Google, Microsoft, Yahoo, Facebook and Digg are starting to participate in data portability related activities by joining in with the Data Portability Work Group (DPWG). Sites like Google and Facebook are taking steps towards unified profiling through initiatives like Friends Connect and Connect. This is just the beginning - there is a lot to discover yet. What is common to all these efforts is the need for a unified profiling system and cross folksonomy data sharing mechanisms.

The advantages of unified profiling and cross folksonomy data sharing mechanisms, for context oriented systems include but are not limited to: a unified user experience across different sites, easy information access for service providing agents like recommender systems and end-user applications, increased recommender productivity due to less time required to search user related information (such as user interests), better planning of retrieval strategies and more accurate evaluation, better equipped

exchange of user information across different social networks and above all meaningful personalization.

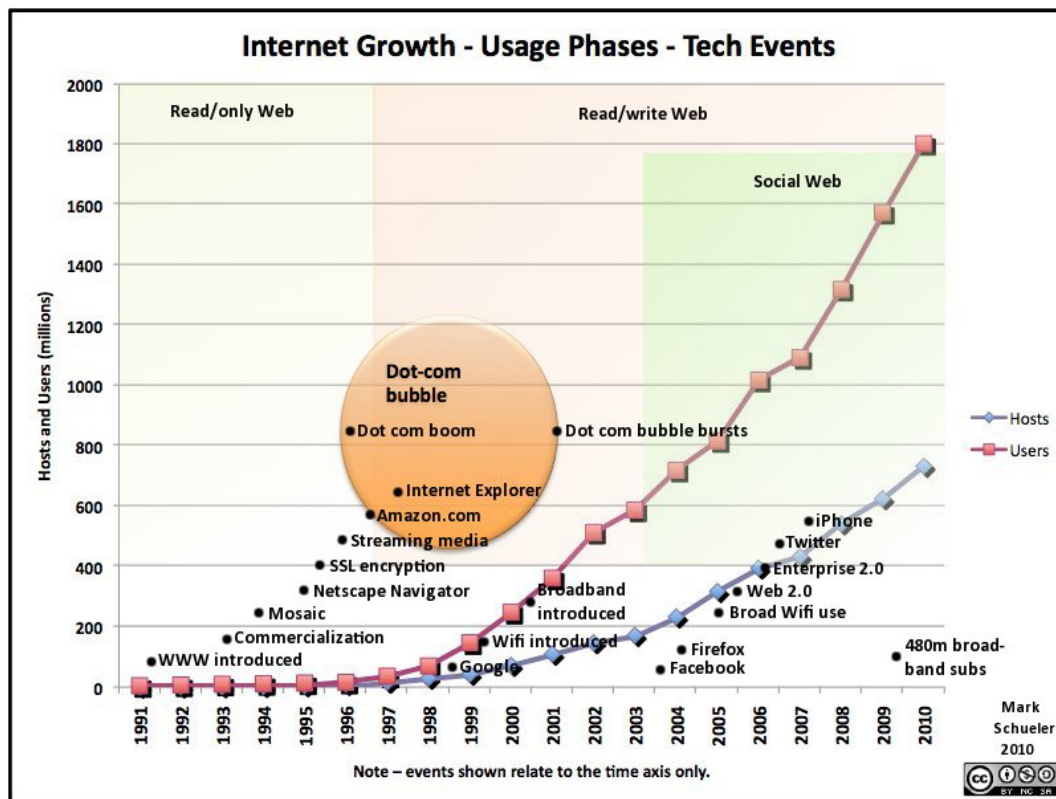


Figure 2.3.2: The Growth of the Web.

Figure 2.3.2 (Schueler, 2010) gives an overview of the growth of Internet usage over the years, but also gives an insight in the technological advancements that led to its popularity and success. As it can be seen with the provision and increase in use of Web 2.0 and SW technologies by the host-websites a drastic increase in the usage of Web and especially SNS is observed.

## 2.4 Social Networks

Social network study is not a new phenomenon in the literature; it has been an important research theme in social science for a long time. As more and more data becomes available due to the Web 2.0 revolution, social networks are gradually being identified as a type of mining resource for all sorts of commercial and research purposes. In this section we review the development of social network sites (SNSs) and discuss the privacy, reputation and trust in them.

#### *2.4.1 History of Social Network Sites*

Computer-based Social Networks have been around for more than two decades though only recently they gained popularity. The idea of using Computer Networks for facilitating computer mediated social interactions goes back to 1979 (Hiltz and Turoff, 1978) who predicted that in the future computerised conferencing will become as common as the telephone. And the computing systems will remove time and distance and will create an environment for ease of sharing thoughts and keeping and making relations. Their vision had most of the things that are a part of today's social networks, by proposing the idea of the first online community, which consisted of synchronous communication (today's live chat), asynchronous communication (messages and discussion boards) and customised news (today's news feeds).

However, early social networking on the World Wide Web began in the form of generalised online communities like the Geocities and Theglobe.com launched in 1995. In these early networks users had to communicate with other users on the site either by using email or other offline methods. That was not the social networking model which is currently being used today. In the late 1990s the user profile became an important component of online social networking sites. Most of the sites started providing facilities like friends lists and search for users with similar interests and by 1997 a new generation of social networks was evolving. The first recognisable social networking site released in 1997 was Sixdegrees.com that we can call the true ancestor of today's social network and the pioneer of the new generation networks. Since the launch of Sixdegrees.com, the Web has witnessed an enormous growth of social networks.

Social networking sites today can be defined as “Web-based services that allow individuals to (1) construct a public or semi-public profile within a bounded system, (2) articulate a list of other users with whom they share a connection and (3) view and traverse their list of connections and those made by others within the system” (Boyd and Ellison, 2007). Figure 2.4.1 illustrates the launch dates of major SNSs and dates when community sites re-launched with SNS features.



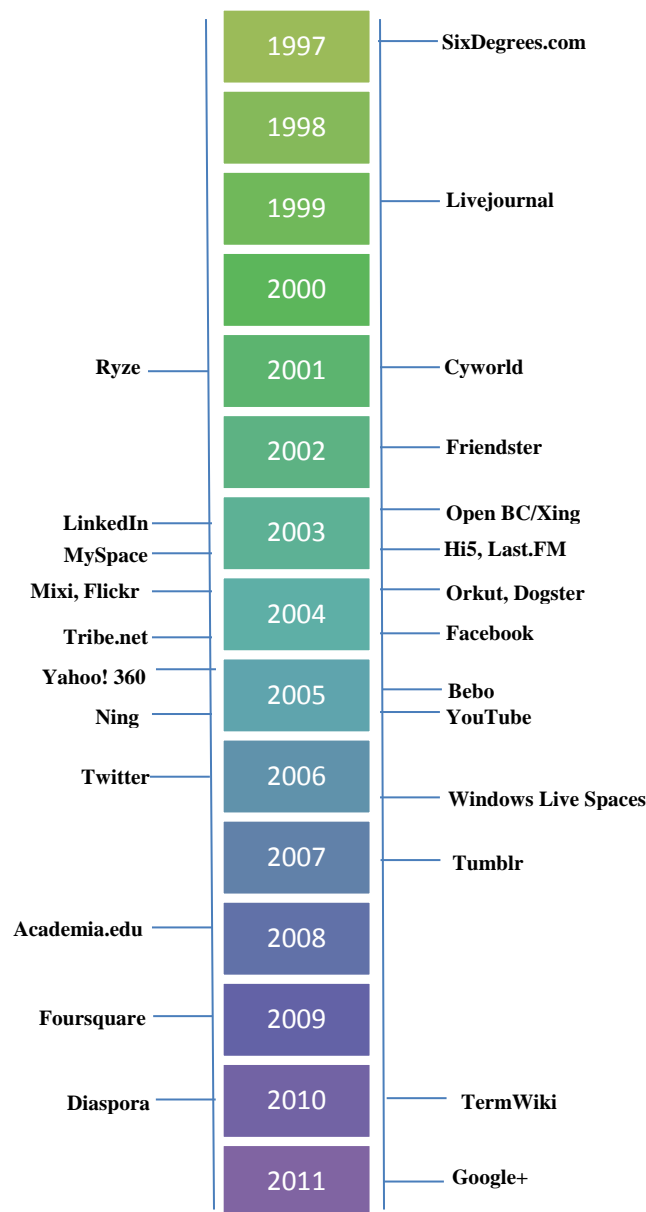


Figure 2.4.1: Timeline of the launch dates of many major SNSs and dates when community sites re-launched with SNS features. Modified from Boyd and Ellison (2007) and Updated.

### *SNS Simply Ahead of Time*

Inspired by the social theory of six degrees of separation, the site Sixdegrees.com was created in 1997. It was promoted as a tool that helped people connect with their friends. The methodology to find a friend was fairly primitive, where people were searched for by name and email address without any details about their profiles. Though a social network, it did not provide services such as blogging and photo sharing, which are integral parts of today's SNSs. Thus there was little to do on the

site after registration. This is probably because the relevant Web technologies were not available or mature at that time. As commented by the founder Andrew Weinreich, “the site was simply ahead of its time”. Due to lack of funding and inability to establish a successful online advertising model consequently after nearly four years of operation, Sixdegree.com, the first social network site, with more than 3 million members, closed down at the end of 2000. However, the Website inspired further development and improvement of SNSs in the following years.

### ***Beginning of a New Era***

The interest in developing social networking sites in the early 2000s geared by Sixdegree.com remained strong, even after the closure of the site due to the dot com recession. Among the emerging sites were Cyworld, Ryze, Hub Culture and Friendster to name a few. Cyworld became the first company to profit from the purchase of online goods that is the purchase of non-physical objects to be used in online communities and online gaming. In 2006 80% of Cyworld’s Korean income was generated through the sale of virtual goods. Ryze was designed to target business professionals, particularly new entrepreneurs. Influenced by the success of Ryze, *Friendster* emerged in 2002, as a social complement to Ryze, competing against the growing number of online dating sites. To register with the site, users were required to create a profile with answers to questions about personal information such as age, occupation, marital status, general interests, music, books, films and television shows. However, unlike most dating sites of the day, which generally introduced strangers to users, Friendster was seeking to introduce friends of friends to users. Users could navigate the social network within four degrees of their personal network. However in June 2011 Friendster shifted from being a social networking site to a social entertainment site.

### ***The SNS as we know them today***

2003 saw a boom in the SNS industry. Venture capital was pouring into the SNS industry. It became obvious that there were huge business opportunities in social network sites. The major Internet players in the industry came to embrace and adopt SNSs due to their huge popularity and commercial success. Google launched *Orkut* in 2004. *Yahoo! 360* was established in 2005. Microsoft introduced its social network platform, *Windows Live Spaces*. With many more small Websites embracing social

network technologies, the SNSs kept growing at a furious pace. MySpace was launched in 2003 to compete against sites like Friendster and Xanga. Gradually it grew its user base, and between 2005 and 2008 it became the most visited site in the world. In June 2006 MySpace surpassed Google as the most visited site in the United States. MySpace was overtaken by Facebook in the unique worldwide visitors in 2008. Facebook started in 2004, as a community site for university students in the US. In 2005, *Facebook* expanded to include high school students, professionals and finally the general public. Facebook promotes itself as “a social utility that helps people communicate more efficiently with their friends, family and co-workers. The company develops technologies that facilitate the sharing of information through the social graph, the digital mapping of people's real-world social connections. Anyone can sign up for Facebook and interact with the people they know in a trusted environment.”(Facebook Factsheet, 2010).

Facebook is one of the most-trafficked sites in the world and has had to build infrastructure to support this rapid growth. It is the largest user in the world of memcached, an open source caching system and has one of the largest MySQL database clusters. The site was mainly written in PHP until the engineering team at Facebook developed a way to programmatically transform PHP source code into C++ to gain performance benefits. Facebook has built a lightweight but powerful multi-language RPC framework that seamlessly integrates infrastructure services written in several languages, running on any platform. Its custom-built search engine serving is entirely in-memory and distributed and handles millions of queries daily. In 2011 Google launched Google+ which is regarded as a strong competitor to facebook. Although Google+ is in its early stages, below is a comparison of the services provided by the two social networking sites to their users. With the launch of Google+, Facebook and Google compete with each other in the following areas as listed in Figure 2.4.1. We have yet to see how these services will compare to each other in the long run, but Table 2.4.1 gives an initial comparison.

facebook.		Google	
facebook.com/games		GAMES	plus.google.com/games
facebook.com/messages		MESSAGING / EMAIL	gmail.com
facebook.com/messages		INSTANT MESSAGING	talk.google.com
facebook.com/videocalling		VIDEO CALLING	talk.google.com
facebook.com/media/photos		PHOTOS	plus.google.com/photos
facebook.com/media/videos		VIDEOS	youtube.com
facebook.com/events		CALENDAR	calendar.google.com

Figure 2.4.1: Facebook vs. Google+ Competing technologies and Services (extracted from veracode.com ‘Google vs. Facebook on Privacy and Security’).

Table 2.4.1: Facebook vs. Google+ initial comparison of the competing features

Competing Features	Google+	Facebook
<b>Google Circles vs. Facebook Friends list</b>	<p>Group of friends you organise by topic</p> <p>The <i>drag &amp; drop</i> option makes it easy to manage groups</p> <p>You can pick and choose different groups while sharing content</p> <p>You cannot exclude groups or friends from getting your updates</p>	<p>Allows to group friends by topic</p> <p>Managing the friends list is not as easy at Google circles. Because there is no real life categorization of connections into friends, co-workers, family etc., were as Google+ provides such categorization.</p> <p>Can choose specific users and groups while sharing content</p> <p>Can exclude certain people or groups from getting your updates</p>
<b>Google+ Stream vs. Facebook News Feed</b>	<p>Looks similar to Facebook’s news feed</p> <p>User can share their photos, videos, links or location for friends and update status. And rate or <i>plus</i> them.</p> <p>Sharing is quicker and user can see who else is able to view their updates.</p>	<p>Unlike Google+ user cannot share their photos, videos, instantly from mobile devices, however can do it via <i>upload e-mail</i>.</p> <p>Can ask questions or poll on facebook wall although such a service is currently not available with Google+</p>
	<p>Excellent feature allows face to face meet up with up to 10 friends/users</p> <p>Chat can be private or</p>	Cannot support group chat.

<b>Google+ Hangout vs. Facebook Video Calling</b>	public  Allows you to watch YouTube videos with other people.  Can be used for group discussion by turning off the video and audio options	
<b>Google+ Location vs. Facebook Places</b>	Adding location to your post is very easy  From the Google+ Android app you can share your location and get updates about the nearby friends	Cannot add location to the feed from the Facebook wall.  Check in places to get updates from your friends nearby is possible on facebook for mobiles.

#### 2.4.2 Privacy Issues

*Privacy awareness:* Privacy has been a concern with social networks since the beginning. In the early days of SNSs, users were usually ignorant of the privacy settings, as observed by Gross and Acquisti (2005) that most of the SNS users did not change their default privacy settings, but rather they manage their privacy concerns by monitoring and limiting the information they share on the SNS. Other research suggests that some users of the SNS are unaware of the visibility of their profile and terms and conditions that apply to them. In contrast Patil and Lai (2005) discovered that despite knowledge of the data exposed the users may not do anything to protect themselves by modifying privacy settings.

*Unauthorised access* is another issue that raises privacy concerns. As Rosenblum (2007) reported and discussed, unauthorised access to user information by third parties is an issue with SNSs.

*More unsecure than other means of communication:* Stutzman (2006) found that social networks pose a more personal and complete disclosure of identity information as compared to traditional methods of communication and information sharing.

*Spamming:* Disclosure of personal information usually attracts spamming and phishing. Zinman and Donath (2007) found that it is more difficult to detect spams in

SNSs than in emails because uninvited messages no longer mean unwanted in social network sites.

SNS is a fast evolving field, and over time SNS have been questioned on the authority of the relations and connections that exist amongst the SNS users. Wilson, et al. (2009) have questioned if the social links of SNSs are valid indicators of real user interaction. They proposed the interaction graph as a more accurate representation of meaningful peer connectivity on social networks. The Facebook Data Team has published a blog about the analysis of the social relationships on Facebook, entitled “Maintained Relationships on Facebook”. They found that on Facebook the number of the mutual relationships, where mutual communications take place between two parties, is far less than that of the maintained relationships, where the users had clicked on another’s News Feed story or visited their profiles more than twice. Our privacy and security measures are explained in Chapter 6 section 6.2.1.

#### *2.4.3 Reputation and Trust*

There are several benefits that the Web can get from the existing social network sites and one of them is the development of reputation and trust of the user in. On the other hand, there are several internet activities like spamming, malware and spyware that pose a threat and need special consideration.

In order to resolve this issue, the technique of governance through peer production has been used by social network sites for which different models regarding reputation and trust are proposed. In this regard, a study by Golbeck and Hendler proposed the use of algorithms that gather the user’s trust relations from other users that are indirectly connected with them through the network (Golbeck and Hendler, 2006). Other techniques discussed by Huynh, et al. (2006) include interaction trust, role-based trust, witness reputation and certified reputation.

A survey by Dwyer, et al. (2007) on Facebook and MySpace users’ experience shows that Facebook members have plenty of trust in both Facebook and its members and similarly MySpace members describe this site as a great tool to meet a new person, which gives a sign of their trust.

Social network systems can be applied to support knowledge sharing between people. Ermecke, et al. (2009) and Domingos (2005) did their studies on marketing benefits of social Websites and found that these sites are very powerful in marketing with effects like viral marketing and knowledge sharing by trustworthy people in a community.

A proof that social networks invade almost all interactions people do, either in real life or online, is available in the studies of Murnan (2006) for emails, Charnigo and Barnett-Ellis (2007) on academic libraries and Baron (2007) for instant messenger platforms.

A critical view by Snyder, et al. (2006) suggests that social networks need to introduce social contract theory to enforce the rules to perform online activities. Also Backstrom, et al. (2006) point out in their research that how much an individual wishes to join a community is influenced not just by the number of connected friends they have but also upon that pattern of connection among themselves.

## 2.5 Web of Linked Data

LD articulates a method of publishing structured data on the Web, so that it can be interlinked and become more useful. It builds upon standard Web technologies such as HTTP and URIs, but rather than using them to assist human readers to access Web pages, it extends Web pages to share information in a format that can be read automatically by computers. This enables data from different sources to be connected and queried (Bizer, Heath and Berners-Lee, 2009).

In recent years more and more data sources on the Web have started to provide access to their databases through APIs, examples being those of Google, Amazon, eBay, Yahoo and many others. Different APIs have different access and identification mechanisms and they provide data in different formats. Because most sources do not provide globally unique identifiers for their items it is not possible to hyperlink them to other items from other APIs. This result, in limiting the choices a developer has to mash-up data from different sources and thus the capabilities of the data cannot be explored to their fullest. To overcome this fragmentation of available data on the Web Sir Tim Berners-Lee outlines a set of best practices for releasing structured data on the

Web “...for exposing, sharing and connecting pieces of data, information and knowledge on the Semantic Web using URIs and RDF” (Berners-Lee, 2006).

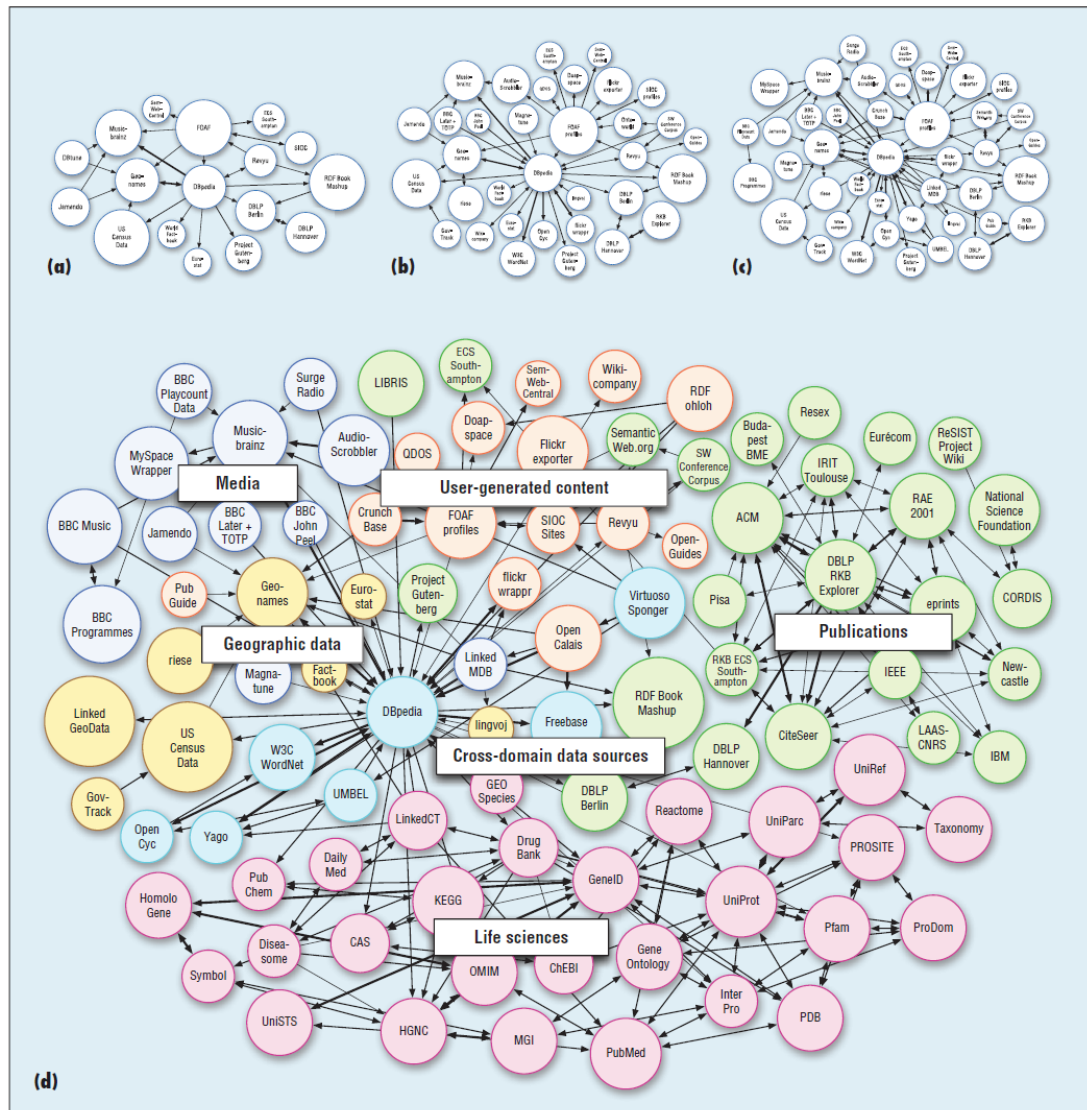
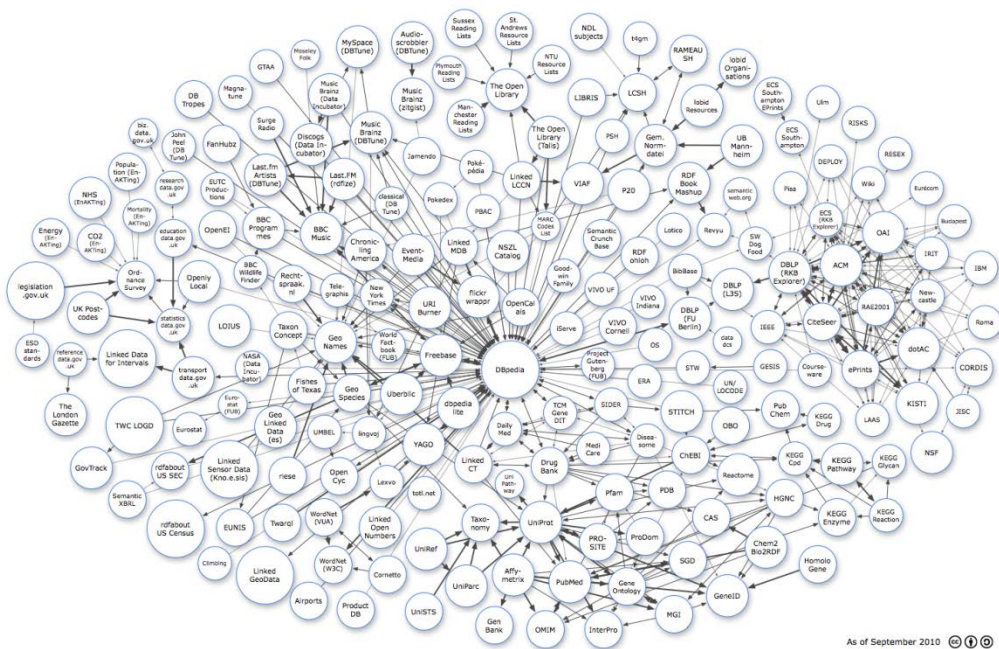
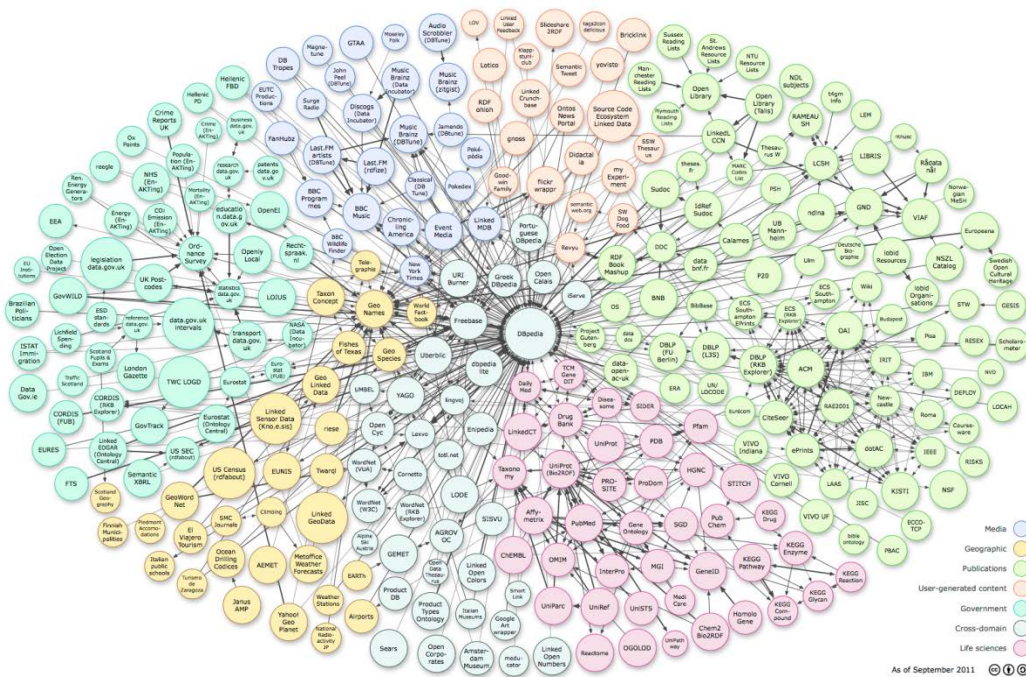


Figure 2.5.1: Growth of Linked Open Data cloud: (a) July 2007, (b) April 2008, (c) September 2008 and (d) July 2009. (Bizer, 2009).





(e)



(f)

Figure 2.5.2: Growth of the Linked Open Data cloud: (e) September 2010, (f) September 2011.

LD opens the potential for data to be examined and explored in new ways and to make new connections between data sources. This ability of LD has been realised by many

people on the Web in the last few years, which is evident from the growth of the LD cloud as can be seen in Figure 2.5.1 and Figure 2.5.2.

To understand the LD concept better, consider the traditional hypertext Web which is considered as the “global information space of interlinked documents” (Bizer, Heath and Berners-Lee, 2009) where a user could simply create a link to another Webpage, even if the user did not have any control of that other Webpage.

Now if we consider the Web as a global data space, what needs to be done to achieve a similar interlinking? Consider there is a database of cities and another of their architectural landmarks. Couldn't we create links from one database directly to the other database? It would be like creating a foreign key from one table in first database to a completely different table in a different database, which the developer has no control of, and thus is unachievable. This is where LD is useful. The same way one could create a link to a Webpage that one has no control of, with LD, links to data residing in other databases on the Web can be made. LD surpasses the physical barriers of machines using a set of LD principles that are described below:

1. Use URIs as names for things: every record in a database will have a URI as its name. This will act as a globally unique primary key (in data base context).
2. Use HTTP URIs so that people can look up those names: typing the URI in a browser should return information about the record that has the URI as its name.
3. When someone looks up a URI, provide useful information using the standards (RDF, SPARQL): Useful information should be provided with the URI in machine readable form RDF so that interesting and useful things can be done.
4. Include links to other URIs so that people will discover more things: internal (URIs pointing to information in the same database) and external links (URIs pointing to information on the Web) could be made for improving information discovery.

The LD initiative is working towards identification of the best practices for publishing, connecting and structuring data on the Web. Key technologies that are helping towards achieving this vision are URIs (a generic way for describing and

identifying concepts in the Web), HTTP (a simple but effective mechanism for retrieving resources on the Web), RDF (a standard way of describing and structuring resources on the Web) and SPARQL (a standard way to query the structured resources). Following is a table of tools for creating publishing and discovering LD on the Web.

Table 2.5.1: Some well-known tools currently available for creating, publishing and discovering LD on the Web

Tools	What they do
<b>For Creating LD</b>	
sqlREST	Exposes relational databases as a REST-style Web Service.
The Silk framework (Volz, et al., 2009)	(helps in link generation) works against local and remote SPARQL endpoints and is designed to be employed in distributed environments without having to replicate datasets locally.
LinQL framework (Hassanzadeh, et al., 2009)	(helps in link generation) Works over relational databases and is designed to be used together with database to RDF mapping tools such as D2R Server or Virtuoso.
<b>For Publishing LD</b>	
D2R Server (Bizer and Cyganiak, 2006)	A tool for publishing non-RDF relational databases as LD on the Web.
Paget and Zitgist	Tools for publishing relational databases as LD.
Virtuoso Universal Server ( <a href="http://www.openlinksw.com/dataspace/dav/wiki/Main/VOSRDF">http://www.openlinksw.com/dataspace/dav/wiki/Main/VOSRDF</a> )	Service for mapping and provision of RDF data via a LD interface and a SPARQL endpoint.
Talis Platform Platform ( <a href="http://www.talis.com/platform/">http://www.talis.com/platform/</a> )	Provides native storage for RDF/LD. Contents of Talis Platform store are accessible via a SPARQL endpoint and a series of REST APIs that adhere to the LD principles.
Tabulator	Tools for publishing relational databases as LD.
Pubby server (Cyganiak and Bizer, 2008)	Rewrites URI requests into SPARQL DESCRIBE queries against the underlying RDF store.
Triplify toolkit (Auer, et al., 2009)	Supports developers in extending existing Web applications with LD front-ends.
SparqPlug (Coetzee, Heath and	Is a service that enables the extraction of

Motta, 2008)	LD from legacy HTML documents on the Web that don't contain RDF data.
OAI2LOD (Haslhofer and Schandl, 2008)	Is a LD wrapper for document servers that support the Open Archives OAI-RMH protocol.
Hakia	Tools for publishing relational databases as LD.
SIOC Exporters ( <a href="http://sioc-project.org/exporters">http://sioc-project.org/exporters</a> )	LD wrappers for several popular blogging engines, content management systems and discussion forums such as WordPress, Drupal and phpBB.
<b>For Discovering</b>	
Zitgist (Semantic Web Browsers)	Is a semantic data viewer that allows to explore sets of RDF data sources on the Web
Tabulator (Semantic Web Browsers)	Allow to explore sets of RDF data sources on the Web
Hakia (Semantic Search Engines)	For publishing and browsing relational databases as LD.
SenseBot	Semantic Search Engines
Falcons (LD search engines)(Cheng and Qu, 2009)	Falcons provide users with the option of searching LD for objects, concepts and documents.
SWSE (LD search engines)(Hogan, et al., 2007)	Keyword-based LD search services oriented towards human users

As more and more data sources are becoming available as LD, the question now is, whether LD is going to be an effective way to bring our cultural heritage resources like libraries, archives and museums together in the open as a fully connected and integrated source of knowledge? And if so, would this eventually help us look beyond what resource the data belongs to but rather what story the data presents as a whole and how it relates to complete the questions a user is asking. Our research explores these questions.

Beyond LD there is a need to embed context to Web-based user data and link this context with interoperable standard ontologies across the various domains. This is a research issue our thesis explores as well. These are issues that extend well beyond the techniques of LD and form the next set of challenges for the Semantic Web. To address this issue we need to understand and recognise the heterogeneity of the Web data, that may vary from Syntactic (being able to handle different data models and formats) to Schematic (being able to understand different data schemata) or Semantic

(being able to unify the different data models, formats and schemata by mapping them using specific criteria and constraints).

## 2.6 Recommender System

Personalization and user-centred adaptability is a hard target to achieve with the World Wide Web, with it being the largest and most diverse database created by mankind. Recommender systems have emerged as an appropriate solution for users to reduce their decision complexity in such information intensive environments. Designing and evaluating such systems however remains an essential challenge for research and practitioners. A critical task, and perhaps most central to an effective recommender system development, is identifying and obtaining user preferences. In our research we are looking at the possibilities of using social-media as a constant mining facility for user interest gathering, for personalizing search and recommendation processes. Below are is overview of Recommender technology, a discussion of the problems common to recommender systems and a study of what research has been done to answer these issues.

### *2.6.1 Recommender Systems, Information Overload and Personalization*

The information overload, where users find an overwhelming number of results that are largely irrelevant to their information needs is a common research problem shared across several research domains such as Information Systems (IS), Recommender Systems (RS) and Human Computer Interaction (HCI). Recommender systems as a solution for information discovery in such situations have been studied for a number of years. RS have been classified in the literature in many different ways (Resnick and Varian, 1997; Schafer, et al., 1999; Terveen and Hill, 2001). Here we classify RS by their underlying method of recommendation.

**Collaborative filtering RS:** Perhaps the most well-known approach in RS is collaborative filtering (CF), which works by recommending the items using similarities of preference amongst users. This approach does not rely upon the content of the item itself but instead depends upon the users to rate items to indicate their preferences and infers preferences similarities by identifying the overlaps of rated

items across users. A typical user profile in a CF recommender system consists of vectors of items and their ratings and is constantly updated as the user interacts with the system. A variety of techniques are applied to calculate the rating of items in the CF domain. Some systems use time based discounting of ratings to account for changes in user interest (Billsus and Pazzani, 2000; Schwab, et al., 2001). Rating can be binary or real-valued describing the preference of the user. Some systems using numeric values are; GroupLens/NetPerceptions (Resnick, et al., 1994), Ringo/Firefly (Shardanand and Maes, 1995), Tapestry (Goldberg, et al., 1992) and Recommender (Hill, et al., 1995).

The greatest advantage of CF based recommender systems is that they work well for domains with subjective choices like movies and music where variation in taste is responsible for a variation in preference. Another significance of CF systems is that they are completely independent of the machine readable representation of the object it is recommending.

### **Content based recommendation RS:**

Content based recommendation is also a very well researched area and one that is often utilised extensively in recommender systems, usually in domains that have extensive textual content like books, news and Website recommenders e.g. the news filtering system NewsWeeder (Lang, 1995). This approach has its roots in Information Retrieval and information Filtering research. The system usually has an item profile comprised of features deduced from the item. The system builds a content based profile of the user based on the weighting of the item features. Thus a typical profile in a content based recommender system consists of the interests of a user which are deduced by the features of the objects the user has rated. A common approach in these systems is to use the content of the items to generate bag of word profiles for users considering their activities and then choose items most relevant to the profiles of the users as recommendations.

Direct feedback from the users can be used to produce the weights reflecting the importance of certain attributes (Joachims, 1997). Some of the popular implementation examples of this approach include the Pandora Radio recommender system that plays music with similar characteristics to that of a song provided by the

user as an initial seed. IMDb (<http://www.imdb.com/>) (Internet Movie Database) and the Jinni search engine (<http://www.jinni.com/>) also utilise context filtering hybrid approaches to recommend movies.

### **Knowledge Base RS:**

Knowledge based (KB) systems use inference upon user's needs and preferences to suggest items. In essence all recommender techniques do some sort of inference. KB systems differ from others because they have functional knowledge to do so i.e., they know how a certain item relates to a certain user's need or preference and therefore can reason upon it to make its recommendations. The user profile in KB systems is a knowledge structure that is a representation of the user's needs and supports the inference mechanism. The KB systems have three types of knowledge requirements to actively perform their task of recommendation; catalogue knowledge (knowledge about items being recommended), functional knowledge (knowledge about mapping the users need to the items that might satisfy their needs) and user knowledge.

### **Demographic RS:**

These recommender systems work by user categorization based on their personal attributes in different classes. The information about the user is generally obtained by a set of questions or a short survey at the beginning. Then users are categorised based on the answers they have provided into different pre-specified classes. Some of the examples of these recommenders include the (Krulwich, 1997) and (Pazzani, 1999) systems that use machine learning to design a classifier using demographic data. In essence the demographic RS make people to people relationships like RS but using different data. An advantage of the demographic technique is that it may not require a history of users' ratings and activities as needed by the CF and the content based RS.

### **Hybrid RS:**

As the name indicates, hybrid recommender systems are systems that combine two or more recommender techniques to achieve better performance with fewer drawbacks than any of the individual ones. The different types of hybrid recommender systems found in the literature are discussed as follows.

**Weighted:** weighted hybrid systems are systems in which the score of the individual item to be recommended is calculated from the results of all of the available recommender techniques. An example of such a system is the P Tango system (Claypool, et al., 1999).

**Switching:** Such recommender systems use pre-defined criteria to switch between the recommender techniques in order to provide better recommendations. An example of such a system is DailyLearner (Billsus and Pazzani, 2000).

**Mixed:** This type of recommenders' present results calculated from several different recommender techniques at the same time. Such systems are more applicable in situations where it is possible to present more recommendations simultaneously. An example of a mixed hybrid recommender is the PTV system (Smyth and Cotter, 2000).

**Feature Combination:** In this technique features from different recommendation data sources are thrown together into a single recommendation algorithm. This is done by treating the information of the first recommender technique as features of the other one.

**Cascade:** the cascade recommender system adopts a step wise process to refine the recommendations. The first recommender technique is employed to find the initial rating. These are then modified and refined through the second recommender technique. An example of a project applying such a technique is EntreeC, which is a restaurant recommender system.

**Feature Augmentation:** In these recommenders, one recommendation technique is applied to obtain the rating or the classification of the item and that rating is then applied in the processing of the next recommendation technique. An example of such a system is GroupLens (Sarwar, et al., 1998).

**Meta-level:** this recommender uses the model generated by the first recommender as in input to the second algorithm. It differs from feature augmentation because here the whole model is fed as an input to the second algorithm instead of making the system learn the model of the first recommender technique to generate features to use as input



in the second system. Example of such an approach used in the literature is LaboUr (Schwab, et al., 2001).

Outside academic research several online vendors and ecommerce sites, including Amazon, eBay and Netflix, have implemented recommender systems to assist their users mostly relying on collaborative filtering techniques. The following table gives an overview of the problems associated with different recommender system technologies.

Table: 2.6.1: Problems with Recommender Systems

Type of Recommender System Algorithm	Problems Associated	Problem Definition	Reference
<b>Common Problems (in general)</b>			
	Cold Start	CF recommenders usually suffer from a <i>cold start</i> problem, in which the system cannot generate accurate recommendations without enough initial information from user.	Different research projects have tried to elevate this problem in different ways; Introducing pseudo users that rate items (Melville, et al., 2002) and neighbourhood based imputation techniques (Su, et al., 2008) are some of the methods. Another commonly proposed solution is to refer to related information such as the textual content of the item to be recommended (Ganu, et al., 2009).

	Shifts and temporal cycles of user interests	User's interest may vary with time.	Lam, W. and Mostafa (2001) investigate the modeling of changes in user interest in information filtering systems
	Potential Bias Affect	In most cases the recommender strategies are designed by the businesses that own the database of which recommendations are to be made and this may result in potential bias.	
	Most similar items are not always good recommendations	Much interesting and related information remains unexplored as they go unidentified by the similarity matrix.	McNee, Riedl and Konstan (2006) discuss that recommendations that are most accurate according to the standard metrics are sometimes not the recommendations that are most useful to users.
	Recommendations made independent of context	An interesting resource is not always a good recommendation. Context plays a vital role but is often ignored.	
<b>Collaborative Filtering RS</b>			
	New User Problem	The phenomenon where the CF algorithm needs a new user to provide their opinion before the system can make any recommendations	Some research done to overcome this problem includes; (McNee, et al., 2003a; McNee, et al., 2003b and

		for the user. It is an effort on the user's part to enter such information.	Rashid, et al., 2002).
	Sparsity Problem	The phenomenon when the user has only rated a small set of items (a common occurrence) and not much information is available for the RS to find overlap between users to recommend items.	Item-based CF is considered as an appropriate solution and is considered to be better in such situations than the user-item rating based matrix. Other solutions are statistically based techniques such as Naïve Bayes and PLSI and various latent analysis techniques. (Huang, et al., 2004; Yu, et al., 2004)
<b>Content Based Recommender Systems</b>			
	Content Limitation in Domain	The phenomenon in non-textual domains when the Content Based Algorithm does not have enough textual data to analyse and give recommendations. It is a common occurrence in the music and movie domains.	Enriching the resource with rich metadata is seen as a possible solution.
	Analysis of quality and taste	The style and quality of the items cannot be determined as they are subjective features. Grammatical analysis can help with the quality to some extent.	Semantic analysis is a possible solution.


	Narrow Content Analysis	This phenomenon occurs when the content based recommender is unable to recommend items which are relevant but differ in content. This is an inherent problem with the Content Based technique as it recommends items based solely on similarity of content.	The solutions proposed here are costly to use. Lexicon and advanced algorithms used to overcome the problem include (Yates and Neto, 1999; Bezzerra and de Carvalho, 2004).
<b>Knowledge based Recommender System</b>			
	The constraint satisfaction problem	This problem arises when no item satisfies the constraints and the logic/ rules set by the knowledge based system.	Bridge and Ferguson (2002) work on the possibility of relaxing rules to overcome this problem.
	Focus on Domain Attributes	Not all domains have rich attributes. In such domain Knowledge Based systems will have difficulty.	




### ***Placing our Research in RS Domain:***

Because most SNs including Facebook have both textual and social information available, key parts of the past work in recommender systems may be applicable to SNs. However, not much research exists on their application and evaluation. As a result it is quite unclear what techniques may be useful and what modifications might be needed to apply them to user data from different SN domains. Our work not only represents the design space for SN (e.g. Facebook) based recommenders but also explores the use of modifications of established techniques from the well-researched recommender systems domain. Another significant difference between our work and those mentioned above is the creation of a user interest profile from pre-existing user data in SNS generate a user interest model thus avoiding the cold start and populating it with related concepts from the open linked-data resources thus making it suitable for

use across various context intensive information domains (such as cultural heritage).  
The following table places our research in the Hybrid-RS Domain.

Table 2.6.1.2: Possible and actual (or proposed) recommendation hybrids, Reproduced and modified from (Burke, 2002).

	Weighted	Mixed	Switching	Feature combination	Cascade	Feature Aug	Meta-level
<b>CF/CN</b>	P-Tango	PTV. ProfBuilder	Daily Learner	(Basu, Hirsh & Cohen 1998)	Fab	Libra	
<b>CF/DM</b>	(Pazzani 1999)						
<b>CF/KB</b>	(Towle & Quinn, 2002)		(Tran & Cohen, 2000)				
<b>CN/CF</b>							Fab, (Condliff, et al., 1999), LaboUr
<b>CN/DM</b>	(Pazzani, 1999)			(Condliff, et al., 1999)			
<b>CN/KB</b>							
<b>DM/CF</b>							
<b>DM/CN</b>							
<b>DM/KB</b>							
<b>KB/CF</b>					EntreeC	GroupLens (1999)	
<b>KB/CN</b>							
<b>KB/DM</b>							

	Redundant
	Not Possible
	Our Resesrch

## 2.7 Cultural Heritage Online

### 2.7.1 *A Study on Personalization in Museum and Tourist Domain*

To analyse the status of provision of personalization in Cultural Heritage projects so far, we have selected around thirty projects from museum and tourist guide systems and evaluated them across three dimensions of personalization implementation (Fan, and Poole, 2006). These systems may or may not focus on personalization in general but do indicate some interesting implications and possibilities. The study is conducted using a comprehensive framework of personalization proposed in literature (Fan and Poole, 2006). Here personalization is considered as a three dimensional implementation choice.

1. The first dimension is about what to personalise. There are a lot of options in a system that can be personalised for a user, mainly the content, functionality, user interface or the channel.
2. The second dimension focuses on who does the personalization. Personalization can be done either by the system itself without active participation of the user (explicit) or it can be conducted with the help of the user (implicit).
3. The last dimension studies to whom the personalization is provided. That is, whether the personalization is directed towards a single person or a group.

Scope of the Study: The systems are chosen from the following research areas; context-aware browsing, semantic interoperability and retrieval, pervasive access, mobile guides and adaptive systems.

The reason behind choosing projects from context-aware browsing, semantic interoperability and retrieval is because most of the work in the CH domain has been done in the interoperability and retrieval domain and that these provide good bases to develop personalised systems.

Projects focusing on pervasive access, mobile guides and adaptive systems are chosen because a part of the study is to analyse the possibility of using personalised access and pervasive access side by side in order to enhance the user's experience.

## **Projects Involved**

There are numerous museum systems facilitating users in various ways. A few of the most appealing ones are discussed here. The first one is the CHIP project (Wang, et al., 2007; Wang, et al., 2008; Rutledge, et al., 2006) whose objective is to use semantic Web technology and adaptive techniques to provide the users with a personalised access to both the museum and its Website. It was designed for the Rijksmuseum and won the bronze award in the Semantic Web Challenge 2007.

The Second project is the Personal Experience with Active Cultural Heritage (PEACH) project Rocchi (2007; 2004). The third project is the winner of the Silver award in the Semantic Web Challenge 2004 and is The Museum Finland project (Hyvönen, et al., 2005). It is a semantic portal for Finnish museums. The main focus of the project is to solve the interoperability problems of heterogeneous museum collections when publishing them on Web. It is quite successful in solving this problem although there is no provision of personalization so far except for a few minor functionalities but it is a promising system in solving the problem of interoperability in heterogeneous museum collections. The fourth project is the Steve Museum (Chun, et al., 2006.), (Trant, 2009) project and is interesting as it studies the possibility of using social book marking systems and tagging data for the creation of a user profile. It presents three different methods of profile creation and visualization and uses an Add-A-Tag Algorithm for profile building. As the user model is the most crucial requirement for personalization this system provides some interesting insights.

The fifth and sixth system, Marble Museum of Carrara by Ciavarella and Paterno (2004) and the Museum AR by Koshizuka and Sakamura (2000) both focus on enabling more natural user interaction with the mobile guide. Marble investigates the use of scan and tilt interaction techniques, the user starts by scanning RFID tags associated with the artworks, and tilt gestures are used to control and navigate the user interface and multimedia information, while the AR although still at its visionary state hopes to accomplish the same by the visitor wearing glasses that augments information about the object in view.

The Museum of Fine arts (Gool, 1999) in Antwerp and the Tokyo University digital museum project 'Point It' uses a camera to take pictures of the artwork and retrieve information about it. Discovery Point (San Francisco Museum of Modern Art, 2001) focuses on promoting social interaction, while the Electronic Guide Book (Fleck, et al., 2002), (Sherry, 2002) one of the most influential and pioneers in investigating the user behaviour towards mobile assistants in museums is a fascinating project that tries to improve the users' experience in the Exploratorium environment. A few other systems like Scott Voice (Woodruff, et al., 2001; Grinter, et al., 2002), Lesar Segall Museum and C Map (Mase, 2002) are also included. All of the above systems in the Museum domain try to improve the user's learning or leisure experience in a museum to some extent. However they miss the most basic needs of the user, the need to personalise and the need to get affiliated. Some popular tourist guide systems are also considered for the study. The most influential projects in this domain are GUIDE and HIPS/Hippie. GUIDE by Cheverst (2000; 2002) which is an intelligent electronic tourist guide developed to provide tourist information about the city of Lancaster. The system focuses upon the issues of flexibility, context-awareness, support for dynamic information and interactive services. It relies on client server architecture. Based on the closest server point the client identifies the approximate location of the user and provides him/her details about the site.

HIPS/Hippie (Hyper-Interaction within physical space) presented by Benelli, et al. (1999), (Oppermann and Specht, 2000) on the other hand is a nomadic system aimed at allowing people to navigate through the physical and the related information space simultaneously, with a minimum gap between the two. HIPS takes into account both an in-door and an out-door scenario of a tourist guide and it can suggest the tourist the most important objects in the surroundings and automatically provide information about them. Being a nomadic system HIPS allows continuous access to information spaces (both the user's personal information space as well as the public information space) independent of specific devices. It thus allows personalization to the visit to the museum, city or any other place of interest according to the user's needs.

Another interesting project is CRUMPET (Creation of user friendly mobile services personalised for tourism) described by (Schmidt-Belz, et al., 2001; Poslad, et al., 2001). It focuses mainly on agent technology. The user can request information and



recommendations about the tourist attractions and the system provides proactive tips about the Point of Interest (POI). The user interests are acquired by tracking user interactions and, thus a history of user context is maintained. The logical user context is based on a domain taxonomy of tourism related services and the probability of user interest in that service. The user context is acquired dynamically but can be accessed by the user. CRUMPET offers adaptation at all levels, i.e. content level, hypertext level and presentation level. All adaptations are performed automatically. The user is given some control over the adaption.

A more recent project in the same domain is the COMPASS2008 (Uszkoreit, 2007) which was designed as a city exploration guide for the Beijing Olympic Games 2008 the main purpose of the system is to assist in language barriers and information access anytime anywhere, as it focuses on cross lingual, multilingual and multi interaction models.

Another example system is the Phone Guide. It deals with pervasive identification of the objects in a museum. The user could load the application in his mobile set and use it. However, no personalization is provided. The KeepUp Recommender System is a hybrid system as it merges collaborative filtering and content based analysis techniques. It is an RSS recommender and it keeps the user up to date with the news that the user is interested in. Cyber Guide (Abowd, et al., 1996) is another system compared, being a mobile tour guide system based on context-aware technology.

Lastly, we consider three prototype tour guide systems: m-To Guide, Sightseeing4u and Gulliver's Genie. m-To Guide (Kamar, 2003) is designed for city travellers and uses GPS technology; it integrates external services and allows transactions such as buying a ticket. It provided preplanning and after-tour support. Sightseeing4u (Baldzer, et al., 2004; Scherp and Boll, 2004) is a personalised city guide and is based on mobileMM4u framework and Niccimon. And Gulliver's Genie (O'Grady and O'Hare, 2004; Hristova and O'Hare, 2003) focuses on artificial intelligence and agents. The user can add personal comments and share information with other tourists. Electronic compass and GPS are used for user location and a dedicated agent manages user context and profiling.

### First Dimension (What is personalised)

As stated earlier the first dimension raises the question of what could be personalised, and this suggests that mainly the content, functionality, user interface or the channel/information access (the media through which information is delivered) can be personalised.

Among the Museum systems CHIP, PEACH, Steve Museum, Museum AR and Museum of FineArts show some degree of content personalization. The interface personalization facility is provided only by CHIP and mSpace. Channel access personalization is provided by almost none.

Among the tour guides Crumppet, COMPASS, Gulliver's Genie, m-To Guide and Sightseeing4u show some kind of content personalization. Functionality and Interface personalization are more common in these systems as compared to the museum guides, but few statistics for personalization of channel access are available..

The following table describes how each of the analysed systems contributes towards personalization.

Table 2.7.1.1: Comparative study of Implementation choices in personalization

System	Who does it?		To Whom?		What?			
	Implicit	Explicit	Individual	Categorical	Content	Functionality	User	Channel/Info
<b>SYSTEMS FOR MUSEUMS</b>								
CHIP	☆	★	★		★		★	
PEACH	☆	☆			☆	☆		
MuseumFinland	⊖		⊖		⊖		⊖	
Steve Museum		★	★		★			
Marble Museum	⊖		⊖		⊖		⊖	
Lesar Segall								
Electronic	⊖		⊖		⊖		⊖	
eChase/mSpace	★						★	
Point it	⊖		⊖		⊖		⊖	
Museum AR	☆				☆			
Museum of Finearts in	☆				☆			
Imogi	★		★			☆		
Discovery point								
Scott Voce	⊖		⊖		⊖		⊖	

Point of	☹		☹		☹		☹	
C-Mape	☆				☆			
<b>TOURIST AND CITY GUIDE SYSTEMS</b>								
Crumpet		★	★		★			
HIPS/Hippie	★	☆	★			☆	☆	☆
GUIDE	☆	☆	★			★		
COMPASS2008	★			★		☆	☆	
Mspace Phone								
Phone Guide	☹		☹		☹		☹	
KeepUp	☹		☹		☹		☹	
COMPASS(2004)	★	☆	★		☆			
Cyber Guide								
CoolTown								
Gullivers Genie		★	★		★			
m-To Guide		★		★	★			
Sightseeing4u	☆	★	★		★			

- ★ Indicates that the particular facility is provided.
- ☆ Indicates that the functionality is provided to some extent or partially provided.
- ☹ Indicates that no personalization is provided at all

Note: Sad face indicates no personalization is provided by the system. While blank space indicates personalization is provided by the system but this particular feature of personalization is not addresses.

## Second Dimension (Who does Personalization)

The second dimension focuses on who does the personalization, .whether it is performed explicitly or implicitly. It was observed that most of the new systems provide explicit personalization, but an initial input (mostly in the form of a question answer dialog and some explicit knowledge during the tour), from the user is always required. However some older systems follow completely implicit personalization procedure. Explicit personalization might be required in situations where there is not much information about the user's current needs/goals and hence the system cannot provide personalization implicitly. Thus in this case the system asks the users explicitly about the information required to personalise the search. While implicit personalization is necessary to reduce the start-up time and assist the uses by making

use of the information that the system already has about the user. A better system will balance the use of both kinds of personalization.

The details of the statistics can be obtained from table 2.7.1.1 and the pie chart in figure 2.7.1.1

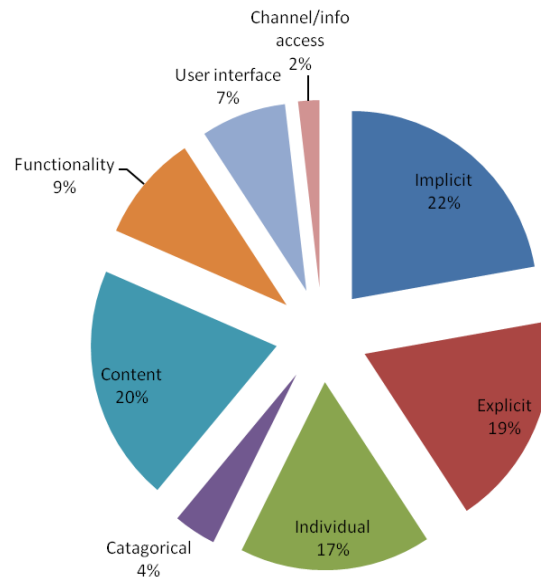


Figure 2.7.1.1: Percentage distribution of implementation choices in personalization among the studied systems

### Third Dimension (To whom is personalization provided)

The last dimension considers, to whom the personalization is provided, that is whether the personalization is directed towards a single person or a group. Table 2.7.1.1 shows that most of the systems studied provide personalization to the individuals rather than the groups.

### Discussion

The study indicates an incline towards systems that focus on applications that know where the user is, what he/she is looking at, what kind of questions the person might ask, and provide the ability to interact with other people and the environment. Not forgetting the fact that none of this can ever be achieved in its true sense without knowing the user, in other words without understanding the interest and needs of the user to provide personalization.

In most of the systems studied, personalization came as a by-product of adaptation, since the main purpose of the system was mostly context aware recommendations i.e., the system was designed to adapt itself to the current context of the user, where context may refer to a variety of factors e.g., users current location, weather, time etc. In other words, the recommender implicitly and inadvertently provides some kind of personalization. There is a strong need to research what kind of personalization is needed for a CH system as the amount and the type are of crucial importance.

### 2.7.2 A Study on Pervasive Access in Museums and Tour Guides

After investigating the state of personalization in the existing CH systems, now the question arises of how this personalization is provided to the user wherever and whenever required. What are the technologies used, how effective are they and what architectural properties make them possible? A selection of the systems from above is chosen for this brief study.

The pervasive computing, combines current network technologies with wireless computing, voice recognition, Internet capability and artificial intelligence, to create an environment where the connectivity of devices is embedded in such a way that the connectivity is unobtrusive and always available. It is important to mention here that not all the aspects of pervasiveness are considered here but only those that are necessary for the provision of a personalised access to the resources.

Table 2.7.2.1: Comparative Study of Pervasive Access in Museum and Tour Guide Systems

System	Architectural Distribution	Awareness Technology	Pervasive/Non-pervasive(Independent)	Mobile/Fixed	Location Identification method.
CHIP	Server based	RFID	Pervasive	Mobile + Fixed	--
PEACH	Server based	IrDA	Pervasive	Mobile + Fixed	Topological
MuseumFinland	Web based	N/A	Independent	Fixed	N/A
Steve Museum	Web based	N/A	Independent	Fixed	N/A
Marble Museum	Information stored in the device itself	IrDA	Pervasive	Mobile	Topological
Lesar Segall Museum	Server based	IrDA	Pervasive	Mobile	Topological

Electronic Guidbook	--	--	--	--	--
eChase/mSpace	--	--	--	--	--
Crumpet	Server based	--	--	Mobile	GPS
HIPS/Hippie	Server based	IrDA	Pervasive	Mobile	Topological
GUIDE	Server based	--	Pervasive	Mobile	Network access point/Network
COMPASS2008	Server based	--	--	Mobile	--
Mspace Phone	--	--	--	--	--
Phone Guide	--	--	--	--	--
KeepUp Recommender	Web based	N/A	Independent	Fixed	N/A
COMPASS(2004)	--	--	--	--	--
Cyber Guide	--	--	--	--	--
Gullivers Genie	Server based but information is cached	--	--	Mobile	GPS + Electronic
m-To Guide	Server based	GPRS	Pervasive	Mobile	GPS
Sightseeing4u	Server based	--	Independent	Mobile	N/A
Museum AR	Server based	IrDA	Pervasive	Mobile	Topological
Museum of Finearts in Antwerp	Server Based	IrDA	Pervasive	Mobile	Topological
Imogi	Information stored in Bluetooth transmitter	Bluetooth	Pervasive	Mobile	Topological
Discovery point	--	--	--	--	--
Scott Voce	Information stored in the device itself	Not required	Independent	Mobile	N/A
Point of Departure	Information stored in the device itself	Not Required	Independent	Mobile + Fixed	N/A
C-Map	Server based	IrDA	Independent	Mobile + Fixed	N/A

The term Topological is used (in the table) where a system specific network topology is used rather than a general/commonly used communication network e.g.; in case of HIPS (1999) a central system represented by a workstation or by a PC acts as Server in the LAN of the considered building or open-space, managing the information system and the interactions with the users. Similarly, in GUIDE (2000) a number of WaveLAN cells, are installed in the city supported by a GUIDE server and portable GUIDE units obtain positioning information by receiving location messages that are transmitted from strategically positioned base stations

## 2.8 Summary of Related work

. Following is a look at the state of the art in some other related research projects.

*Information overload:* Bobillo, et al. (2008) address the problem of information overload by defining ontologies for context as well as domain knowledge. They describe a scenario of outdoor assistance for health care where context dependent information is provided for patient treatment. Kim, et al. (2003) propose a user's interest hierarchy (UIH) for defining a user's general and more specific interests. They suggest that text and phrases from user's bookmarked Web pages can be used to identify his/her general and specific interests.

*Semantic ambiguity:* Kauppinen, et al. (2008) addresses the problem of semantic ambiguity in geographic place names and tries to address this by designing ontologies for places (SUO and SAPO). These ontologies are published at a local server to be utilised as a mash-up service later on in their system (CULTURESAMPO portal).

*User Profiling and SNS:* Unified profiling and tag data portability among different social sites is gradually coming into prominence, in the research community. The credit for this realization and initial efforts is shared among the bloggers as well as the developers of these communities. A plethora of projects are trying to answer these issues, providing interesting results on user information in tags. Studies on social networks indicate that users intend to tag contents they are interested in with descriptive tags that can be used to identify their interest (Li, Guo and Zhao, 2008). Athanasis (2004) shows how tag clouds from multiple social Websites demonstrate a tendency to overlap regardless of the type of folksonomy the website uses. The work also suggests the tendency of profile enrichment through cross-linking of tag clouds. Angeletou (2008) presents FLOR, a mechanism to automatically enrich folksonomy tag-sets with ontological knowledge. Gruber (2008) suggests that true collective intelligence can be achieved by linking user contributed contents and machine gathered data, and that the social Web and the semantic Web should be combined into collective knowledge systems. With this visible possibility the Semantic Web can play a vital part in describing tags and relating them to meaningful concepts in ontologies.

*Ontologies:* Significant efforts have been made to describe ontologies for tags, taggers and tagging systems. SICO ontology (Bojars, et al., 2008) aims to define the main concepts that are needed to describe a social community on the semantic Web. The aim is to view a person's entire contributions on the social Web. The FOAF (Friend of

a Friend) (Brickley and Miller, 2005) ontology helps describe people. The SKOS (Simple knowledge organization system) (Miles, et al., 2004) is a model for sharing and linking knowledge organization systems like thesauri, taxonomies and vocabularies via the semantic Web. The SCOT (Social semantic clouds of tag) (Kim, et al., 2007) ontology presents a model for the main properties and concepts required to represent a tagging activity on the semantic Web. Similarly the MOAT (Meaning Of A Tag) (Passant and Laublet, 2008) ontology, as the name indicates, is a collaborative framework to help Web 2.0 users give meanings to their tags in a machine readable format. Promising work here is the Google Social Graph API. The API returns Web addresses of public pages and publically declared connections between them which help identify and track various Web identities of a user and thus assist in the collection of tag clouds related to an individual.

We find it the right time to make an effort to utilise semantic Web standards and ontologies to enrich the data from unified profiling systems in order to make it useful in semantic search and recommender systems.

*Related Projects and Research:* Some of the projects that have tried similar approaches include (Sinclair, Lewis and Martinez, 2007) that proposes an automated link service that uses Wikipedia as its link-base for linking data with concepts, and (Specia, Motta, 2007) that focuses on determining relations among tags in social networks to form clusters based on concepts from ontologies. This work suggests that by exploiting Wikipedia, Word Net, Google and semantic Web ontologies, meaningful relations can be identified amongst tags.

Li, et al. (2008) suggest a mechanism to identify the social interest based on user-generated tags. They propose an Internet Social Interest Discovery system (ISID) which works on the principle of clustering users and their saved URIs based on common frequently-occurring tags. These clusters identify the topics of social interest.

Iturrioz, et al. (2007) suggest a transition from desktop to Web where more and more users are keeping their resources on the Web; like pictures in flickr, bookmarks in del.icio.us and so on. Despite significant ease and advantages, this has resulted in the fragmentation of user resources and therefore a global view of resources is needed.



Their work is a loosely coupled federated tag management system that provides a uniform view of tagged resources across different Web 2.0 sites.

Szomszor, et al. (2008) present a way of determining an individual across flickr and del.icio.us by assessing his/her tags, filtering them utilising Google Search and Word Net and finally forming a FOAF based user profile. Perhaps the most related project to our work is by Cantador, et al. (2008). It builds upon the tag filtering mechanism developed in (Szomszor, et al., 2008) and moves further to design ontological profiles for tag users. This is done by matching tags with ontological concepts. Users are not passive consumers of content in in Social Networks (SN). They are often content producers as well as consumers. We investigate the potential of the content generated by the user to understand the user better and answer the following questions in this process.

- Can SN help users find interesting contents on CH Websites?
- How can user interest information obtained from SN lead to better recommender design?

## 2.9 Conclusions

This chapter presents the literature review for this thesis. Important background information and the developments over the years in the field of personalization in the Semantic Web, and the Social Web, the emergence of the Web of linked data, Recommender Systems and the provision of personalization in Cultural Heritage is discussed. Section 2.2 we discussed the semantic Web in the light of past research in the field of personalization in research areas like adaptive hypermedia and Web mining. This section also frames the relation between semantic Web research and the social Web movement and highlights the possibilities that mutual research in these areas can reap.

Section 2.4 summarises the history of the Social Networking System (SNS), and discusses the evolution of Facebook as one of the most popular SNS on the Web. Here we discuss some of the major issues like privacy and trust in using SNS. This review is important here because Facebook plays a major role in testing two of our research hypotheses. The first hypothesis about, *using social data as the recourse for gathering*

*user interest and second hypothesis that, states a mechanism for designing automated query formulation through the use of SN data.*

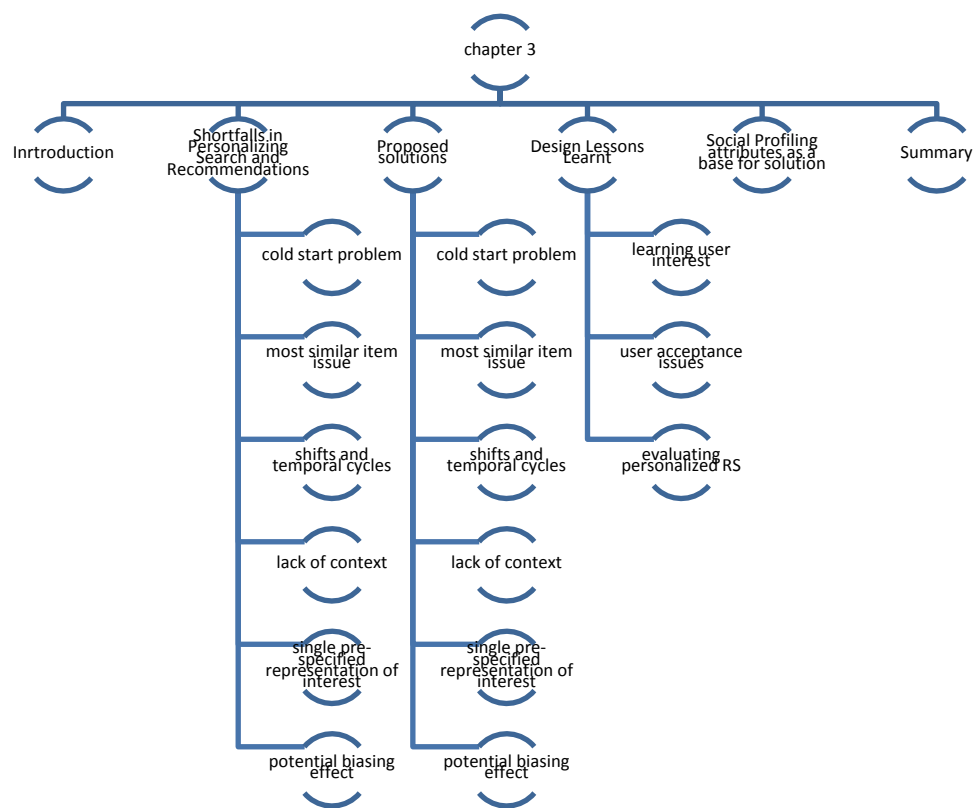
In Section 2.5 this chapter discusses the linked data initiative and how it fits into the picture of this research.

Next in Section 2.6 we give a detailed review of the Recommender System domain, the problems that are inherent with the use of different RS techniques and related work in solving these problems. Further on in Section 2.7 this chapter presents a literature review of the online cultural heritage community including the major research and commercial projects in the domain. We further discussed the state of the art in personalization and pervasive access in museums and tour guide systems and present a short survey of systems and their comparison. This chapter presented the state of the art in projects that aim to answer similar questions as our research is looking into.



## Chapter 3

# Problem Definition and Solution Design Issues



Chapter 3 topic hierarchy

### 3.1 Introduction

The importance of personalised access to cultural heritage online and the need for recommender systems in such similarly information overloaded domains, where automated suggestion of items of interest to users have become increasingly essential, has been discussed in previous chapters. Personalization is highly valued in such areas to overcome the evident problem of information overload. Today's search systems are advanced and sophisticated enough to conduct the process of locating user tailored search results thanks to extensive research and joint contributions from the fields of Information Retrieval, Recommender Systems, Adaptive Hypermedia and User Modeling. These solutions, however, are not free of serious pitfall like the very well-known 'cold start' problem. This Indicates that there is a considerable need for systems that improve dynamic profiling and interest gathering, this will help the system to better understand the requirements of a particular user and will aid in finding and integration of more suitable results and presentation of information in a personalised manner.

*In this chapter it is argued that:*

*The Cold start problem is a common problem in personalised recommender systems and its root cause is lack of user interest information and or ways of capturing it.*

*The problem of finding and updating user interest information unobtrusively and dynamically while relating it to appropriate concepts to suggest relevant information resources is still not solved.* This is essential for any recommendation or filtering process to occur in a personalised environment. This suggest, that users can still benefit from personalised systems that gather user interest information unobtrusively to support grouping or filtering of information according to the user's current interests thus assisting them to avoid the information overload problem, or to help them discover new information "hidden" in an information resource.

This chapter discusses in Section 3.2 the various shortfalls in personalizing search and recommendations such as; the cold start problem, the fact that the most relevant items are not always good recommendations, shifts and temporal cycles of user interests, lack of context or recommendations made independent of context, lack of updating user interest or the case when only the items specified in one pre-specified

representation are considered and the potential bias effect. Section 3.3 gives suggestions to solve these issues. Section 3.4 presents a brief discussion on the lessons learnt regarding the provision of personalization. Section 3.5 places the user's social network profiling data as the basis of the solution and Section 3.6 concludes the chapter.

The chapter therefore discusses the problems to be solved and possible directions to solve these problems, including why social network data can form the base of the solution.

A review of the process, methods and design features that permit deliberation when dealing with information personalization tasks is also undertaken to look into considerations that should be taken for the final solution.

## 3.2 Shortfalls in Personalizing Recommendations

### 3.2.1 *The 'cold start' Problem*

*Cold start* is a problem common across different types of recommender systems as is evident from the literature review in the previous chapter. It is the state of the system in which the system cannot generate accurate recommendations without enough initial information from the user.

In Collaborative Filtering (CF) recommenders the similarity between user profiles is assessed to predict ratings for unseen items. The shortcoming of the cold start problem in the CF method occurs due to the technique's assumption that active users will respond positively to unseen items rated highly by similar users (Pennock, et al., 2000). As most users are not inclined to rate previously seen items, only a few items will receive ratings resulting in scarcity of data needed to produce recommendations and hence the cold start problem. The similarity metrics generated in such cases are not sensitive enough to distinguish between users, particularly new users of the system (Schein, et al., 2002). Hence, the most highly rated items from anyone are recommended. The reverse effect is also present, i.e. a newly imported item cannot be

recommended until it receives sufficient ratings. Different research projects have tried to elevate this problem in different ways; introducing pseudo users that rate items (Melville, et al., 2002) and neighbourhood based imputation techniques (Su, et al., 2008) are some of the methods. In statistics, imputation is the substitution of some value for missing data. Many imputation techniques are available e.g. hot-deck imputation or imputation is also done with the use of machine learning classifiers (Rahman, Davis, 2012). Another commonly proposed solution is to refer to related information such as the textual content of the item to be recommended Ganu, et al. (2009; 2012) and Tsatsou, et al. (2009).

The Content based recommender systems work by matching the characteristics of the items to be suggested with the relevant features in the user profile. This requires the system to model sufficient details of the user's preferences and interests through preference elicitation into the user's profile. This is achieved either by querying the user for required information (explicitly) or by observing the behaviour of the user over time (implicitly). In both cases the cold start problem persists as the user has to dedicate a certain amount of time and effort using and helping a system in a 'not so helpful' state working towards the construction of their own profile before the system could provide them with any intelligent recommendations.

### *3.2.2 The most similar items are not always good recommendations*

Recommender systems have shown great potential to help users find interesting and relevant items from within a large information space. Most research up to this point has focused on improving the accuracy of recommender systems. We believe that not only has this narrow focus been misguided, but has even been detrimental to the field. Consider for example Content based filtering (CBF) approaches that basically index the items of possible interest in terms of a set of automatically derived descriptive features, then unseen items with similar attributes to those rated highly are recommended. A drawback of this method of recommendation is that it recommends items interchangeable with those that have previously received high ratings, by virtue of its focus on the items' features, ignoring potential user requirements. As such, for a system to be able to avoid such issues, equivalence between items, in a particular user context, needs to be evaluated.

Another possibility in such cases is to return results that a user is already familiar with, indicating that even the recommendations that are most accurate according to the standard metrics are sometimes not the recommendations that are most useful to users. Imagine you are using an art recommender system. Suppose all of the recommendations it gives to you are for artwork you have already viewed. Unfortunately, this is quite possible in current recommender systems. In the standard methodology, the art recommender is penalised for recommending new artwork instead of artwork the users have already viewed. Current accuracy metrics, such as Mean Absolute Error (MAE) as analysed by Herlocker (2004), measure recommender algorithm performance by comparing the algorithm's prediction against a user's rating of an item. The most commonly used methodology with these metrics is the leave-n-out approach (Breese, 1998) where a percentage of the dataset is withheld from the recommender and used as test data. In essence, the system reward an art recommender for recommending artwork a user has already viewed, instead of rewarding it for finding new artwork for the user.

McNee, et al. (2006), propose that the recommender community should move beyond the conventional accuracy metrics and their associated experimental methodologies, and embrace user-centric directions for recommending and evaluating recommender systems.

### *3.2.3 Shifts and temporal cycles of user interests*

Most conventional RS architectures do not model for shifts of the user's interest over time, since all ratings provided by a user have an equal bearing on producing recommendations.

### *3.2.4 Recommendations made independently of context*

An interesting resource does not necessarily make a good recommendation every time. Typically, recommender algorithms do a good job of identifying resources that are similar (or relevant) to those already consumed by users. In most cases however, they fail to capture the reason the user is seeking recommendations. As such, the



recommended resources may fail to fulfil the user's recommendation need, while being interesting at the same time.

#### *3.2.5 Only items described in one pre-specified representation are considered*

Since the focus in RS applications has been to enable organisations to suggest appropriate items from their catalogue to customers, not much effort has been put into learning user preferences based on the items they already have in their possession, regardless of their origin. However, a good sales assistant in a clothing shop will first look at what the customer is wearing before making suggestions.

#### *3.2.6 Potential biasing effects*

Following on from the previous point, the fact that the provider of the recommendation service is typically the vendor of the resources available for recommendation introduces new considerations. Since the vendor stands to profit from the users of the RS and resources have varying profit margins, it is highly conceivable that they will introduce bias towards producing recommendations that if consumed would maximise profit for the vendor. Further, it can be expected that in situations where the system cannot make any recommendations with high confidence, popular items are recommended in the hope of a sale. Both these phenomena have been observed and are seen as diverting the focus away from satisfying user needs.

### 3.3 Proposed Solutions

#### *3.3.1 The 'cold start' problem*

The Cold start problem is considered to be as the main problem to be solved in the context of the proposed framework. The solution is achieved by ensuring that users are not assigned empty profiles upon registration, but rather carry with them the information that reflects their current interests across multiple domains. Of course, if users have not created any information prior to subscribing to the system (or have chosen to not disclose any) the problem persists. However, such behaviour would somewhat defeat the point of seeking personalised recommendations. The user interest

information is automatically gathered from user's social networking account and is dynamically updated. To avoid initial and constant updating efforts required in making the user profile up to date, linking the user interest model with the users SNs' profile is proposed as a solution. Similarly, resources for recommendation are imported only if the user profile indicates a strong interest in resources of that particular type.

### 3.3.2 *The most similar items are not always good recommendations*

The fact that the framework requires each resource to be mapped to a unique set of terms in the universal vocabulary (DBpedia) provides a mechanism for identifying interchangeable resources. Such resources are expected to have identical descriptions using terms from the universal vocabulary and can therefore be merged. This mechanism calculates the equivalence amongst items which is the basic solution for the item similarity issues. This will increase the 'Findability' of previously hidden yet related information aiding in new knowledge discovery and elevating the problem to some extent.

The interest filter proposed in the system (though mainly designed to solve the problem of shifts and temporal cycles of user interests) also has a side advantage that it helps elevate the 'similarity of item' problem, by ensuring that the recommended results are always from a set of highly weighted resources across a set of resource types (best representing users current interest) rather than from a single type of resource.

In our proposed model the query results are presented in order of relevance, but to avoid the most *similar items are not always good recommendation* phenomenon, we suggest a novel approach of filtering (see chapter 6, section 6.2.7) the results thus obtained with the top five resource types that the system has calculated to be most related to the user current interest profile. This we believe will bring variety to the results without losing relevance to the user interests.

In addition the automatic upgrading of the interest profile each time a user logs a new interest in their SN ensures a dynamic interest profile that forms the core for the

recommender system keeping sure that the current interest is always considered alongside the relevance matrix while handling a user search or query string.

### *3.3.3 Shifts and temporal cycles of user interests*

The rich representations used by the profiling component of the framework enable the segmentation of user profiles to be based on contextual attributes such as location. The dynamic nature of user profiling allows recording of shifts and temporal cycles of user interest as the user profile is constantly updated with the user's SN data. Information created or accessed by a user during a specific time interval can easily be selected from their profile log, although there is no need to consider this information in the current framework. The temporal information is only captured for recording log times in this research. As such, shifts of interest cycles and location can be accommodated within the framework by considering only the most recent (subjective to a threshold) elements of a profile in order to make recommendations, which is inherent in the use of SN data.

### *3.3.4 Recommendations made independently of context*

Object attributes alone are not adequate for making recommendation. The framework developed here offers a solution for automatically determining which aspects of a user interest profile are relevant to the context of a particular query. This is achieved this by providing a novel search tool which overlays the user current interest rating with the context of the user's query to produce results explicitly selected to reflect a particular context.

### *3.3.5 Only interests described in one pre-specified representation are considered*

The notion of an exhaustive index of the resources to be recommended does not exist in this framework. Instead, the search results are filtered through a user interest matrix (as explained in chapter 6). Any resource type that the framework finds related (through semantic annotation and ranking) to a user interest, that is, where the user has implicitly expressed interest in it as part of their profile, regardless of their origin, is considered. By adopting this mechanism, as such, the effects of problems associated

with the inadequacy of user profiles to represent a wide range of user interests are expected to be less severe.

#### *3.3.6 Potential biasing effects*

The framework shifts the emphasis from the business aspect of the recommender system to satisfying user needs in deploying RS technologies. By basing the end results on a standalone user preference/interest calculation mechanism independent of the end data resources, it becomes harder to spuriously insert an arbitrary recommendation. Hence a recommendation cannot be added to the list from the system manager. Moreover, to influence the system to recommend the said resource over others, one would also have to gain control over the universal representations of resources and the semantic connections between their descriptive terms in the universal vocabulary. Furthermore, since the SN data are simply seen as platforms indicating the preferences of their members, there is no guarantee of what objects will be selected for a user, on the bases of extracted SN data, as a recommended resource.

### 3.4 Design Lessons learnt regarding Personalization

#### *3.4.1 Learning the user interest*

Robust user profile construction is a significant design part of personalization systems. The work in the area of Information systems, especially content based information retrieval, traditionally uses techniques like frequent patterns and click history (Wang and Zeng, 2011), which though useful are not flexible enough to represent user interests, term weighting schemes, implicit interaction data (Melville, et al., 2002), statistical language modeling (Song and Croft, 1999; Zhai, 2008) and long term search history. Although these techniques have proved useful in the information system domain and deliver useful prospects for a user profile to develop once a user has invested sufficient time using the system, they do not compensate for the initial lack of user interest information also known as ‘new user problem’.

### *3.4.2 User Acceptance issues in Recommender System*

As shown in previous work by McNee (2006) and Ziegler (2005), user satisfaction does not always correlate with high recommender accuracy. There are many other factors important to users that need to be considered. New users have different needs from experienced users in a recommender. New users may benefit from an algorithm which generates highly rateable items, as they need to establish trust and rapport with a recommender before taking advantage of the recommendations it offers. Related work by Rashid (2002) shows that the choice of algorithm used for new users greatly affects the user's experience and the accuracy of the recommendations the system could generate for them. The literature review in this area also suggests that differences in language and cultural background influence user satisfaction (Torres, 2004). A recommender in a user's native language was greatly preferred to one in another language, even if the items themselves recommended were in the other language. (E.g. an Urdu-based research paper recommender recommending papers written in English).

### *3.4.3 Evaluating Personalization Recommender Systems*

There are many different aspects to the recommendation process which current accuracy metrics do not measure. As they do not take into consideration the unpredictability of the human interest factor and the fact that it is a very abstract variable to evaluate. McNee, et al. (2006) discuss and review an additional three such factors that are not captured by traditional evaluation metrics namely: the similarity of recommendation lists, recommendation serendipity i.e. an unexpected recommendation and the importance of user needs and expectations in a recommender. They further review how current methodologies fail for each aspect and provide suggestions for improvement.

## **3.5 Social Profiling Attributes as a Base for Solution**

Social networks have long been an important research theme in social science. As data about large-scale networks are increasingly available, social networks are gradually identified as a type of mining resource for all sorts of commercial and research

purposes. To support this argument we presented recent research on SNSs in Chapter 2 section 2.4. These included research on online social capital, privacy issues, personalization, reputation and trust.

Take the example of Facebook which is a popular online SN. It is said that “Every fourteenth person in the world is a facebook user” (Facebook, 2010). This statement is enough to prove the social impact of Facebook on the world. Facebook provides a service to users to publish and share their personal, private and public information and experiences on the Internet, as many times and at any time of the day. This makes this site an intensely personalised database of more than half a billion active users (750 million) in the world. This makes it most suitable to use as an interest mining resource for user profile creation. The feasibility and usefulness of the data thus collected is evaluated through two preliminary evaluations described in Chapter 4 and is further established by the evaluations of the Cheri Recommender and Search Systems in chapter 7.

Because most SNs, including Facebook, have both textual and social information available, key parts of past work in recommender systems may be applicable to them. However, little research exists on their application and its evaluation methods. As a result it is unclear what techniques may be useful and what modifications might be needed to apply them to user data from different SN domains.

Our work not only represents the design space for SN (e.g. facebook) based recommenders but also explores the lessons learnt from established techniques from the well-researched recommender systems domain. Another significant difference between our work and those mentioned above is the creation of user interest profile from pre-existing user data in SN to generate a user interest model. and populating it with related concepts from the open linked-data resources thus making it suitable for use across various context intensive information domains (such as cultural heritage) while addressing the cold start and related problems explained in section 3.2 of this chapter.

### 3.6 Conclusion

In this chapter it is argued that the cold start problem is a common problem in personalised recommender systems and its root cause is lack of user interest information and/or ways of capturing it. The problem of finding and updating user interest information unobtrusively and dynamically while relating them with appropriate concepts to suggest relevant information resources is still not solved.

In section 3.2 the various shortfalls in personalizing search and recommendations are discussed including the cold start problem, the fact that the most relevant item is not always a good recommendation, shifts and temporal cycles of user interests, lack of context or recommendations made independent of context, lack of updating the user interest or the case when only the items specified in one pre-specified representation are considered and the potential bias effect.

Section 3.3 gives suggestions to solve these issues. It is suggested that the cold start problem can be avoided by ensuring that users are not assigned empty profiles upon registration. The user interest information is automatically gathered from the user's social networking account and is dynamically updated. To avoid the initial and constant updating efforts required in making the user profile, linking the user interest model with the users SNs' profile is proposed as a solution. With the understanding that the most similar items are not always good recommendations a mechanism for identifying interchangeable resources through a shared universal vocabulary (DBpedia) is suggested. An interest filtering method is also proposed in the system, which though mainly designed to solve the problem of shifts and temporal cycles of user interests, also has a side advantage that helps alleviate the 'similarity of item' problem, by ensuring that the recommended results are always from a set of highly weighted resources across a set of resource types best representing the user's current interest rather than from a single type of resource (see section 6.2.7 for more details). The rich representations used by the profiling component of the framework enable the segmentation of user profiles to be based on contextual attributes such as location. The dynamic nature of user profiling allows recording of shifts and temporal cycles of user interest as the user profile is constantly updated with the users SN data. In the issues with recommendations made independently of context it is realised that the

object/resource attributes alone are not adequate for representing the context of a recommendation.

The framework offers a solution for automatically determining which aspects of a user interest profile are relevant to the context of a particular query. This is achieved by providing a novel search tool which overlays the user's current interest rating with the context of the user query to produce results explicitly selected to reflect a particular context. The search results are filtered through a user interest matrix, the interest matrix comprises of weight concepts related to users interest, extracted from the users SNS profile. Any resource type that the framework finds related (through semantic annotation and ranking) to a user interest, that is, the user has implicitly expressed interest in as part of their profile, regardless of their origin, is considered. By adopting this mechanism, as such, the effects of problems associated with the inadequacy of user profiles to represent a wide range of user interests are expected to be less severe.

The framework shifts the emphasis to satisfying user needs in deploying RS technologies. By introducing a standalone user preference/interest calculation mechanism independent of the end data resources, it becomes harder to spuriously insert an arbitrary recommendation. Moreover, to influence the system to recommend the said resource over others, one would also have to obtain control over the universal representations of resources and the semantic connections between their descriptive terms in the universal vocabulary. Furthermore, since the SN data are simply seen as platforms indicating the preferences of their members, the objects selected for a user as a recommendation are based solely on the users own preferences explicitly identified by them on their SNS pages.

Section 3.4 presents a brief discussion on the lessons learnt regarding the provision of personalization, covering topics like user interest capturing, user acceptance issues in recommender systems and evaluating personalised recommender systems.

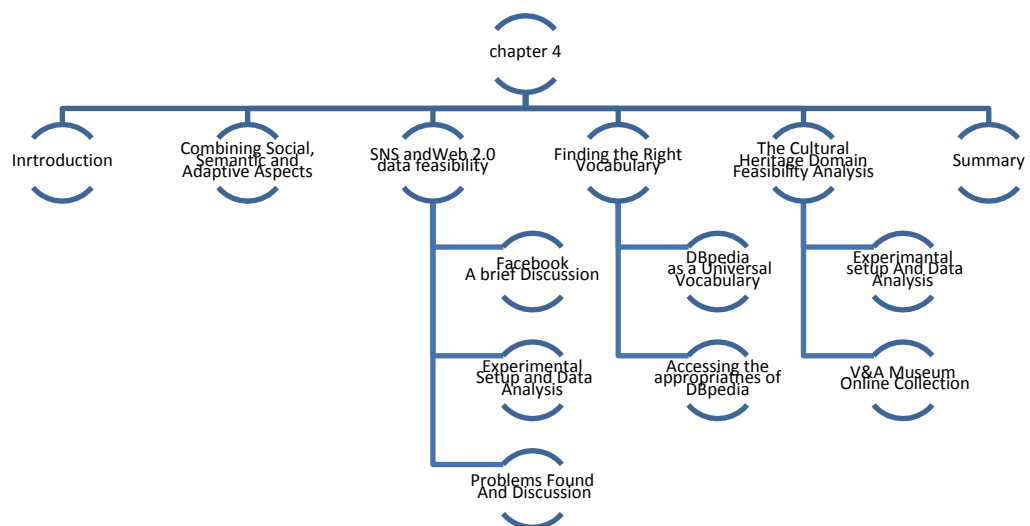
Section 3.5 places the user's social network profiling data as the basic part of the solution. The feasibility of using the SN is further discussed and tested in Chapter 4. This chapter therefore refines the problems to be solved and discusses possible



directions to solve these problems, suggesting that social network data can form the bases of these solutions.

## Chapter 4

# Feasibility Studies and Preliminary Tests



Chapter 4 topic hierarchy

This chapter describes the initial work we undertook to justify the use of Social Networks/ Web 2.0 data as a source of gathering user interest information, more precisely a discussion on Facebook as the Social Network of choice for this type of user data mining. With regard to identifying the problems with the Web 2.0 user generated data; this chapter resolves the issues by using appropriate vocabularies. The study to identify the most appropriate vocabulary for the type of data we were extracting is discussed next. The justification and reasoning of using a universal vocabulary rather than a more specific vocabulary is studied and discussed. Finally finding and testing the feasibility of the data corpora that we used to generate the knowledge base for artwork recommendations is presented. This section describes the Victoria & Albert Museum in London's online collection database and the freely available linked open data on the Web as our data sources of choice. This initial work provides a base for the choices we made to eventually create the idea of the Cheri system (chapter 5) and to implement it as a recommender and search system (see chapter 6) using the semantic Web and LD standards.

## 4.1 Combining Social, Semantic and Adaptive Aspects

Personalization and user-centred adaptability is a hard target to achieve with the World Wide Web, as it is the largest and most diverse database created by mankind. In our research we are looking at the possibilities of using social-media as a constant mining facility for user interest gathering, for personalizing search and recommendation processes. Such a task holds challenges of integrating user data from various sources, resolving conflicts and abstracting from and reasoning about the data thus obtained (Noor and Martinez, 2009). It involves overcoming syntactic and semantic heterogeneity of distributed user models in order to achieve usable and interoperable user interests. In this chapter we establish a case for the use of social data for personalization (using cultural heritage as our experimental domain) and discuss some results. Such an approach we believe can assist in reasoning and personalization activities in search and recommender systems especially in context intensive and task oriented scenarios.

Social media today is increasingly becoming an essential contributor for keeping us informed in our social, personal and professional lives. Every morning with the first

cup of coffee and our delightful groovy gadgets, most of us tiptoe into an information world filled with possibilities. With the ease of use, availability and least skill requirement as its selling points, most of us have grabbed the social Web with both hands; though we are still struggling to keep a balance of privacy, openness, safety and freedom. Everyday tremendous amounts of information about people, places, events and products are generated using different social websites. And each piece of information thus created, viewed, shared, commented upon or ignored has the capacity to take us one step farther in understanding the users better and as a result help improving their experience on the Web.

The question is how to use this data? Two of the major problems here are; establishing credibility for the use of social data in reasoning and query refinement tasks; and overcoming the sparse semantic structure of social data. Our research is exploring ways to solve these two problems. The approach we take towards solving these issues is by annotation and mapping of social data to existing yet richer vocabularies and semantic sources, while retaining their original context. This chapter presents in Section 4.3.2 some of the evaluation on social profile data from Facebook, in order to establish a case for the use of social data. This is followed by an experiment in Section 4.4.2 to demonstrate the usability of this data for search refinement through personalization in the cultural heritage domain. The idea will eventually evolve into (by chapter 5) a cultural heritage oriented, term refinement and query optimization model which we believe will further improve the search and recommendation process by providing a suitable basis for improved query optimization.

## 4.2 Social Networks/Web 2.0 Data Feasibility

This section presents a discussion on the feasibility of using Social Network data for identifying user interests. Facebook is chosen as a case study. Section 4.2.1 starts by an overview of the critiques facebook has faced over the years and concludes with a discussion on why despite these critiques Facebook is still the most suitable Social Media example for use in our research. Section 4.2.2 lists the problems found in Web 2.0 users' data and forms the basis for discussion in the next section.

Data has taken the lime-light in today's networked economy. We live in a world built and wrapped in data and as we interact with it, we create more. Data is not only the Web's core resource, it is fast becoming a ubiquitous commodity, fuel and necessity in real time, that follows and surrounds us nearly everywhere we go. Data is a resource which is renewable and reusable and Social Networks along with other Web 2.0 applications are becoming one of the greatest producers of user data. We shape the world of data collectively with each purchase, search, status update, news feed and tweet many times each day. We believe that this user data should be harvested as a renewable resource to facilitate the use of data elsewhere on the Web and in our day to day lives in a meaningful manner.

#### *4.2.1 A brief discussion about Facebook*

The Web of data is scaling to nearly incomprehensible size and power. In order to understand where and how user data is created and spreads in a real world environment, we wish to examine a setting where a large set of population of individuals frequently exchange information with their peers. Facebook is the most widely used social networking service in the world, with over 800 million people using the service each month. For example in the United States, 54% of adult internet users are on facebook (Hampton, et al., 2011). Those American users on average maintain 48% of their real world contacts on the social site (Hampton, et al., 2011) and many of these users regularly exchange news items with their contacts (Kossinets and Watts, 2009). Thus Facebook represents a broad online population of individuals, whose online personal networks reflect their real world connections, making it an ideal environment to study user interest dynamics and information contagion, which is a useful phenomenon considering one of the aims of this research is to introduce the general Web users to cultural heritage related information according to their interests in a seamlessly unobtrusive yet pervasive manner.

We begin by a discussion on the controversies and critiques surrounding Facebook since it was first introduced in 2004.

### *Controversies surrounding Facebook:*

This is a brief look at criticism on Facebook regarding technical issues (note this does not include the non-technical concerns like litigation and third party responses to facebook, or inappropriate content controversies and other such matters). Table 4.2.1.1 gives a brief history of the issues raised and what amends if any done have been made to overcome these issues by Facebook.

Table 4.2.1.1: Facebook critique issues and amend made: a brief history.

Year	Critique	Issues Raised and Amends
14 <sup>th</sup> Decem ber 2005	Facebook started being criticised on its <b>use as a means of surveillance and data mining</b> . The early case of Data mining by private individuals unaffiliated with facebook came to public light. Two Massachusetts Institute of Technology (MIT) students were able to download, using an automated script, over 70,000 Facebook profiles as part of a research project on facebook privacy (Jones, Harvey, Soltren and Hiram, 2005)	Since then, Facebook has boosted security protection for users, responding: "We've built numerous defences to combat phishing and malware, including complex automated systems that work behind the scenes to detect and flag Facebook accounts that are likely to be compromised (based on anomalous activity like lots of messages sent in a short period of time, or messages with links that are known to be bad)". (Fred Wolens, Retrieved December 2010)
July 2007	<b>External privacy threats:</b> Concerns started that <b>Facebook could be used to violate privacy rules or create a worm</b> when Adrienne Felt, an undergraduate student at the University of Virginia, discovered a cross-site scripting (XSS) hole in the facebook Platform that could inject JavaScript into profiles. (Felt, 2007)	This hole took facebook two and a half weeks to fix (Felt et al., 2008)
August 2007	<b>Internal Privacy threat: Source code leak.</b> A configuration problem on facebook server caused the PHP code, to be displayed instead of the Web page the code should have generated (the code was responsible for Facebook's dynamically generated home and search) (Cubrilovic, 2007)	This raised concerns about how secure the private data on the site was.
Septem ber 2007	<b>Opening user data to the general Web.</b> Facebook drew a fresh round of criticism after it began allowing non-	In the following months Facebook's privacy settings, however, allowed users to block

	members to search for Facebook users using general search engines, with the intent of opening limited "public profiles" up to search engines such as Google (BBC Online, 7 <sup>th</sup> September 2007).	their profiles from search engines if they wished to do so.
November 2007	A system called <b>Beacon</b> was introduced by Facebook which allowed third party websites to add a script by facebook on their site and use it to send information about the actions of Facebook users on their site back to Facebook.	This raised serious privacy concerns.
29 <sup>th</sup> November 2007	In the initial launch of beacon the information was automatically published.	It was changed to require confirmation before publishing later.
1 <sup>st</sup> December 2007	A security engineer at CA Inc. claimed in the blog post that Facebook collected data from the affiliate sites even when the consumer opted out and even when not logged into the facebook site.	Beacon was discontinued September 2009.
February 2008	<b>Data Ownership concerns:</b> a <i>New York Times</i> article in February 2008 pointed out that facebook does not actually provide a mechanism for users to close their accounts and thus raised the concern that private user data would remain indefinitely on Facebook's servers (Aspan, 2008) Facebook had allowed users to deactivate their accounts but not actually remove account content from its servers. If a user wanted their data removed the user had to clear their own accounts by manually deleting all of the content including wall posts, friends and groups. Still there were concern that emails and other private user data remains indefinitely on Facebook's servers (Aspan, 2008)	Facebook subsequently began allowing users to permanently delete their accounts. Facebook's Privacy Policy now states: <i>"When you delete an account, it is permanently deleted from Facebook"</i> ("Facebook Privacy Policy", <a href="http://www.facebook.com/about/privacy/">http://www.facebook.com/about/privacy/</a> , December, 2010.)
February 2009	<b>'Terms of use' controversies</b> of Facebook started when a blogger Chris Walters claimed that the changes Facebook made to its terms of use on 4 <sup>th</sup> of February 2009 gave facebook the right to "Do anything they want	In order to calm criticism, Facebook returned to its original terms of use. However, on February 17, 2009, Zuckerberg wrote in his blog, that although Facebook reverted to its original

	<p>with your content. Forever." (Walters, 2009)</p> <p>In January 2011 Electronic Privacy Information Center (EPIC) filed a complaint claiming that Facebook's new policy of sharing users' home address and mobile phone information with third-party developers were "misleading and fail to provide users clear and privacy protections", particularly for children under age 18. (Complaint, Request for Investigation, Injunction and Other Relief, <a href="http://epic.org/privacy/inrefacebook/EPIC_Facebook_Supp.pdf">http://epic.org/privacy/inrefacebook/EPIC_Facebook_Supp.pdf</a>, January 2011)</p> <p>Facebook temporarily suspended implementation of its policy in February 2011, but the following month announced it was "actively considering" reinstating the 3rd party policy.</p>	<p>terms of use, it was in the process of developing new terms in order to address the paradox. A new voting system with two new additions to facebook: the Facebook Principles and the Statement of Rights and Responsibilities (Zuckerberg, M., Thursday, 26 February 2009) was introduced next. Both additions allow users to vote on changes to the terms of use before they are officially released. However, the new terms of use released were harshly criticised in a report stating that the democratic process surrounding the new terms is disingenuous and significant problems remain in the new terms. The report was endorsed by the Open Rights Group. (<a href="http://www.openrightsgroup.org/">http://www.openrightsgroup.org/</a>).</p>
October 2009	<p><b>Controversial News Feeds:</b> Facebook launcher news feed and mini feeds on 5<sup>th</sup> September 2006.</p> <p>In October 2009, Facebook redesigned the news feed with the focus on popular content, determined by an algorithm based on interest in that story; including the number of times an item is liked or commented on. Live Feed would display all recent stories from a large number of a user's friends.</p> <p>In December 2009 Facebook removed the privacy controls for the news feeds and the mini feeds. This change made it impossible for the user to control what activities are published on their walls (and consequently to the public news feed).</p>	<p>The changes brought in October 2009 met immediately with criticism from users. Users did not like the amount of information that was coming at them and people couldn't select what they saw.</p> <p>Since December 2009, people could publish anything they wanted. This allowed people to post things that could target certain groups of people or abuse other users through other means.</p>
November 2009	<p><b>Reducing users' privacy and pushing users to remove privacy protections</b> In November 2009,</p>	<p>Facebook has since re-included an option to hide friend's lists from being viewable; however,</p>



	<p>Facebook introduced a new privacy policy. This new policy made certain information, including <i>lists of friends</i>, as publically available information, with no privacy settings; it was previously possible to keep access to this information restricted. Due to this change, the users who had set their <i>list of friends</i> as private were forced to make it public without even being informed and the option to make it private again was removed (Rom Cartridge - What is Facebook?", <a href="http://romcartridge.blogspot.co.uk/2010/01/what-is-facebook.html">http://romcartridge.blogspot.co.uk/2010/01/what-is-facebook.html</a> ) This was protested about by many people and privacy organizations such as the Electronic Frontier Foundation( EFF) (<a href="https://www.eff.org/">https://www.eff.org/</a> ) and the American Civil Liberties Union(<a href="http://www.aclu.org/">http://www.aclu.org/</a> ). The change was described as <i>Facebook's Great Betrayal</i> (Bankston, 2009)</p> <p>Mark Zuckerberg, Facebook CEO, had hundreds of personal photos and his events calendar exposed in the transition.</p>	<p>this preference is no longer listed with other privacy settings and the former ability to hide the friends list from selected people among one's own friends is no longer possible (McCarthy, C., December 11, 2009).</p> <p>Defending the changes, founder Mark Zuckerberg said "we decided that these would be the social norms now and we just went for it".</p>
	<p><b>Cooperation with government search requests without reasonable suspicion.</b> An article by Junichi Semitsu published in Peace Law Review stated that "even when the government lacks reasonable suspicion of criminal activity and the user opts for the strictest privacy controls, Facebook users still cannot expect federal law to stop their 'private' content and communications from being used against them." (Semitsu, 2011)</p>	<p>Facebook policy reads: "We may also share information when we have a good faith belief it is necessary to prevent fraud or other illegal activity, to prevent imminent bodily harm, or to protect ourselves and you from people violating our Statement of Rights and Responsibilities. This may include sharing information with other companies, lawyers, courts or other government entities."(Semitsu, 2011)</p> <p>Since Congress has failed to meaningfully amend the Electronic Communications Privacy Act to protect most communications on social networking sites such as Facebook and since the Supreme</p>

		<p>Court has largely refused to recognise a Fourth Amendment privacy right to information shared with a third party, there is no federal statutory or constitutional right that prevents the government from issuing requests that amount to fishing expeditions and there is no Facebook privacy policy that forbids the company from handing over private user information that suggests any illegal activity. The term fishing expedition is defined as a legal proceeding mainly for the purpose of interrogating an adversary, or of examining his or her property and documents, in order to gain useful information (for information).</p>
August 2011	<p><b>Accessing facebook- kept user records:</b> The group ‘europe-v-facebook.org’ made access requests at Facebook Ireland and received up to 1.200 pages of data per person in 57 data categories that Facebook was holding about them, including data that was previously removed by the users (<a href="http://www.europe-v-facebook.org/removed_content.pdf">http://www.europe-v-facebook.org/removed_content.pdf</a> ). Despite the amount of information given, the group claimed that Facebook did not give them all of its data. Some of the information not included was likes, data about the new face recognition function, data about third party websites that use social plugins visited by users and information about uploaded videos.</p>	<p>This implies that a user cannot request for or get a copy of all the information Facebook is keeping about them. The DPC is investigating the issue and this investigation by the DPC might become one of the most severe investigations into Facebook’s privacy practice in the past years.</p>
	<p><b>Interoperability and Data Portability issues:</b> The inability of users to export their social graph in an open standard format contributes to vendor lock-in and contravenes the principles of data portability (Baker,</p>	

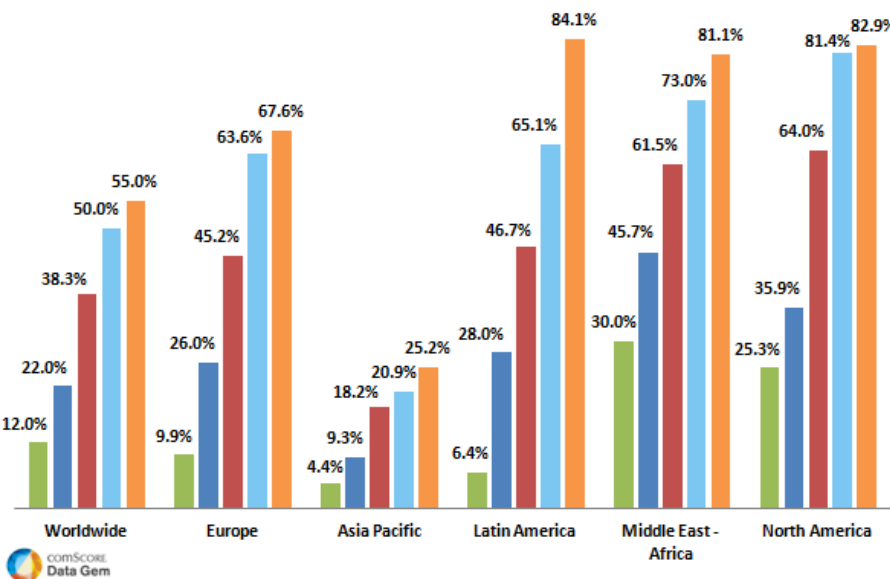
	<p>2008). Automated collection of user information without Facebook's consent and third-party attempts to do so (e.g., Web scraping) have resulted in suspension of accounts (Scoble, 2008), cease and desist letters (Agarwal, 2007) and litigation with one of the third parties, Power.com.</p> <p>Facebook Connect has been criticised for its lack of interoperability with Open ID. "Facebook Connect was developed independently using proprietary code, so Facebook's system and OpenID are not interoperable.... This is a clear threat to the vision of the Open Web, a future when data is freely shared between social websites using open source technologies." (Calore, 2008).</p>	
--	--	--

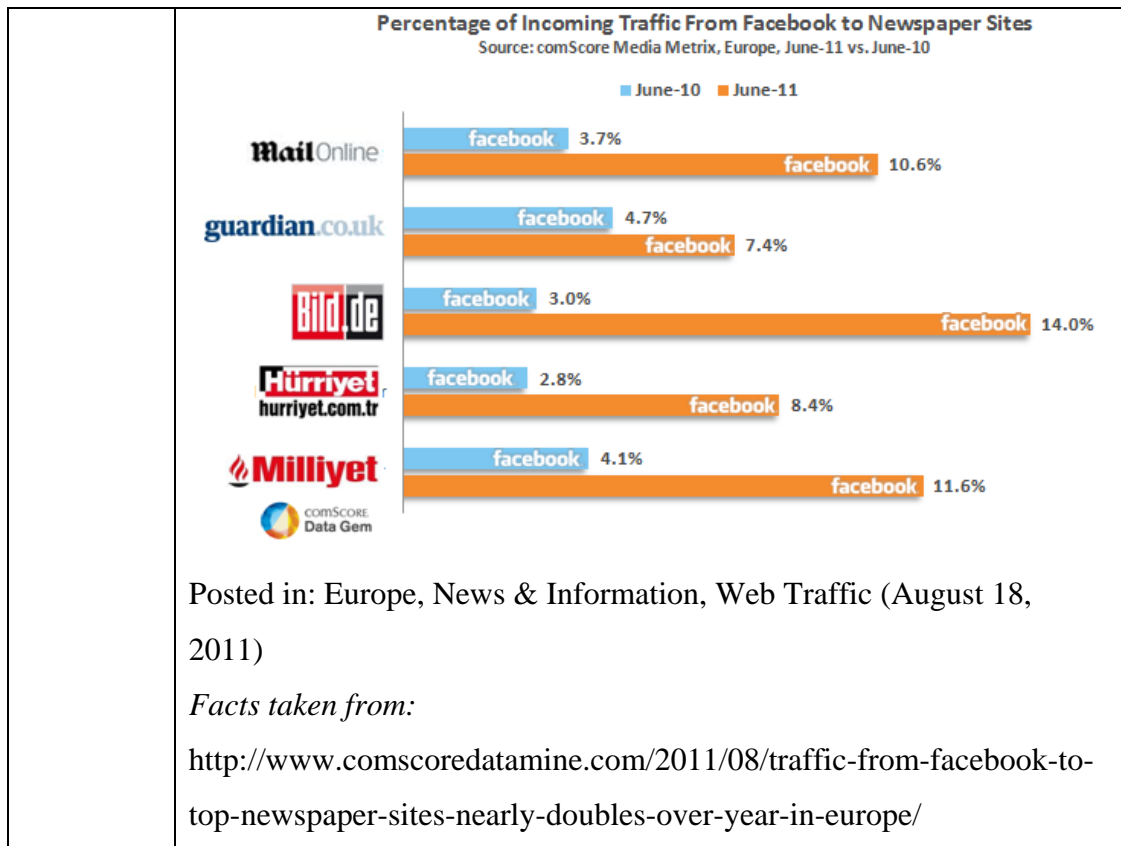
The above table demonstrates the Facebook approach to privacy and data security to be post-active rather than proactive. Facebook policy has always been more of a social experiment where the real time testing performed on user data and the reactions thus obtained are used to mould and remodel the policy and to find yet new ways to open the user data or to accustom users to viewing their data as open. This observation made us realise that as there is not as much of a strong privacy and data protection related policy inherent in Facebook as is needed for our research; we need to take our own measures to ensure that the user data is asked for and used according to the users' permission in our system. Keeping this in mind the decision was made to provide additional measures for the Cheri system to answer the privacy related issues. The Cheri system has the same login system as the facebook but by keeping the user well informed about the type of data that is gathered from their SNS profiles each time and giving the user full opportunity to select and deselect the type of information the user want the system to use helps provide adequate privacy and better user experience. This measure helps answer the threats like opening user data to the web and terms of use controversies. The measures taken by the Cheri system to keep the user data secure and encrypted are mentioned in Chapter 6(Cheri System design) and Chapter 7 (Evaluation and Results). Keeping the user data in FOAF format helps answer the data ownership concerns. As the data remains in the hands of the user instead of being

restricted to the Cheri system and the user is free to use it with any other site that uses FOAF. The reasons why Facebook was chosen despite the inherent privacy concerns and critiques are explained in the following table and discussion.

Table 4.2.1.2: Justification for using Facebook for user data mining

Reason	Brief Discussion/Facts
Popularity	<p>“Facebook has the highest user percentage of all social media sites, With over 500 million users, Facebook is used by 1 in every 13 people on earth, with over 250 million of them (over 50%) who log in every day.”</p> <p><i>Facts taken from:</i></p> <p><a href="http://www.digitalbuzzblog.com/facebook-statistics-stats-facts-2011/">http://www.digitalbuzzblog.com/facebook-statistics-stats-facts-2011/</a></p>
Usage over time	<p>“Over 700 Billion minutes a month are spent on Facebook, 20 million applications are installed per day and over 250 million people interact with Facebook from outside the official website on a monthly basis, across 2 million websites. Over 200 million people access Facebook via their mobile phone. 48% of young people said they now get their news through Facebook. Meanwhile, in just 20 minutes on Facebook over 1 million links are shared, 2 million friend requests are accepted and almost 3 million messages are sent.”</p> <p><i>Facts taken from:</i></p> <p><a href="http://www.digitalbuzzblog.com/facebook-statistics-stats-facts-2011/">http://www.digitalbuzzblog.com/facebook-statistics-stats-facts-2011/</a></p>
Facebook Shows Strong Growth Over Past Five Years	<p>“Facebook’s reach amongst the total internet audience has continued to increase over the past five years across all regions. Globally, Facebook reached 12 % of the internet audience in December 2007 and as of December 2011 the social network reached over half of the internet audience, 55 % (a 43 percentage point rise). With Facebook’s reach increasing across all regions over the past year, it shows that while Facebook already reaches over half of the total internet audience worldwide, its audience is still increasing.”</p>

	<p style="text-align: center;"><b>Facebook's Penetration (%) of Total Internet Audience by Global Region</b> Source: comScore Media Metrix, Age 15+, Home and Work Computer Usage</p> <p style="text-align: center;">■ December 2007   ■ December 2008   ■ December 2009   ■ December 2010   ■ December 2011</p>  <table><thead><tr><th>Region</th><th>December 2007</th><th>December 2008</th><th>December 2009</th><th>December 2010</th><th>December 2011</th></tr></thead><tbody><tr><td>Worldwide</td><td>12.0%</td><td>22.0%</td><td>38.3%</td><td>50.0%</td><td>55.0%</td></tr><tr><td>Europe</td><td>9.9%</td><td>26.0%</td><td>45.2%</td><td>63.6%</td><td>67.6%</td></tr><tr><td>Asia Pacific</td><td>4.4%</td><td>9.3%</td><td>18.2%</td><td>20.9%</td><td>25.2%</td></tr><tr><td>Latin America</td><td>6.4%</td><td>28.0%</td><td>46.7%</td><td>65.1%</td><td>84.1%</td></tr><tr><td>Middle East - Africa</td><td>30.0%</td><td>45.7%</td><td>61.5%</td><td>73.0%</td><td>81.1%</td></tr><tr><td>North America</td><td>25.3%</td><td>35.9%</td><td>64.0%</td><td>81.4%</td><td>82.9%</td></tr></tbody></table> <p><small>comScore Data Gem</small></p> <p><i>Facts taken from:</i> <a href="http://www.comscoredatamine.com/2012/02/facebook-shows-strong-growth-over-past-five-years/">http://www.comscoredatamine.com/2012/02/facebook-shows-strong-growth-over-past-five-years/</a></p>	Region	December 2007	December 2008	December 2009	December 2010	December 2011	Worldwide	12.0%	22.0%	38.3%	50.0%	55.0%	Europe	9.9%	26.0%	45.2%	63.6%	67.6%	Asia Pacific	4.4%	9.3%	18.2%	20.9%	25.2%	Latin America	6.4%	28.0%	46.7%	65.1%	84.1%	Middle East - Africa	30.0%	45.7%	61.5%	73.0%	81.1%	North America	25.3%	35.9%	64.0%	81.4%	82.9%
Region	December 2007	December 2008	December 2009	December 2010	December 2011																																						
Worldwide	12.0%	22.0%	38.3%	50.0%	55.0%																																						
Europe	9.9%	26.0%	45.2%	63.6%	67.6%																																						
Asia Pacific	4.4%	9.3%	18.2%	20.9%	25.2%																																						
Latin America	6.4%	28.0%	46.7%	65.1%	84.1%																																						
Middle East - Africa	30.0%	45.7%	61.5%	73.0%	81.1%																																						
North America	25.3%	35.9%	64.0%	81.4%	82.9%																																						
Traffic from Facebook to Top Newspaper Sites Nearly Doubles Since Last Year in Europe	<p>“In June 2011, Facebook accounted for at least 7.4 % of the traffic going to the top five Newspaper sites in Europe. The German publication Bild.de, which ranks as the third most popular newspaper site in Europe, saw the most incoming traffic from Facebook (14.0 %). It also experienced the most growth over the previous year with an 11-percentage point increase in visitation from Facebook. The British Mail Online, which ranks as the top newspaper site in Europe, saw 10.6 % of visitors coming from Facebook in June 2011 – an increase of 6.9-percentage points over the previous year. Guardian.co.uk, which was the second most popular newspaper site in Europe with 13.5 million unique visitors, received 7.4 % of its visitors from Facebook, growing by 2.7 percentage points over the past year.”</p>																																										



Using Facebook Open Graph:

While conducting the feasibility analysis of SN our data extraction model (previously) used *Facebook Connect* to access publically available user data from facebook (after user authorization), which had a data storage restriction no more than 24 hours. It gave little time for any complex data processing tasks to be performed on the data. As per April 2010 Zuckerberg (Facebook CEO) and Taylor (former *friendfeed* CEO and present Facebook Director of Product) during the F8 Developer conference, introduced three new features to the facebook developers' platform, one of them being "*Open Graph*" that is tailored to be used by businesses and services. Since this introduction our system was shifted to the Open Graph protocol and now operates on this protocol for the extraction of user data.

Alfred Korzybkksi, the father of general semantics, famously remarked "the map is not the territory". However since its introduction the Open Graph by Facebook has been criticised as not being open in the true sense of openness raising remarks like "The missing bit is that Facebook appears to be the only repository of data in this equation - and that makes the whole offering seriously closed". Jeff Jarvis summed it up in a

tweet saying "What we want closed (our data) they want open. What we want open (create and transfer) they want closed." "Open" means that no one "owns" either end of the process, but in the case of Facebook's open graph, Facebook owns the entire process. While this holds true for the bigger competitors in the business, if we look at the Open Graph API specification and policies, it is clear that in principle developers are permitted access, permanent storage and re-purposing (within limits) of the user data acquired, via the Facebook Open Graph APIs. Of course it is very unlikely that Facebook would open its open graph to permit a giant of its own stature to access its entire core asset repository. However it is also hard to prevent Facebook users from accessing their own information through a plethora of existing or new third-party applications. Thus the majority of Facebook's graph data in aggregate through these third-parties applications will be replicated and thus become open over a period of time. The following table gives some pros and cons of using the Facebook Open Graph APIs.

Table 4.2.1.3: Concerns with Open Graph Protocol

Pros of Open Graph	Cons of Open Graph	Cons of Open Graph that Cheri deals with
<p>The open graph can help in <b>synchronization of interests</b> the user has posted on third-party websites he/she has visited. That is, the things that the users show their interest in on a third party website are recorded as interest-nodes on that particular user's facebook open graph. In this way the facebook open graph will eventually show a collective view of the user interests across different websites. This could help in making the browsing experience better.</p> <p>"It will prove to be an</p>	<p>Open Graph Protocol does not support <b>object disambiguation</b>. Although it is simple to use there is no way to disambiguate objects most of the time.</p> <p><b>Incorrect implementation at its launch.</b> Launch partners have not implemented Open Graph Protocol correctly on their sites.</p> <p><b>Lack of mark-up.</b> Facebook does not have the mark-up on its own pages that it asks the world to adopt. The biggest problem with the semantic mark-up is, the ambiguity inherent in the implementation.</p> <p><b>Duplication of data.</b> A growing amount of user</p>	<p><b>Object Disambiguation and lack of mark-up</b> is solved by the Cheri System with the use of DBpedia ontology and vocabularies like WordNet.</p> <p>Cheri also handles <b>Duplication of data</b> problem by generating its own version of user profile in a standardised ontology (FOAF)</p> <p>Generation of a FOAF profile and use of DBpedia categorization allows Cheri interest profile to have <b>secondary attributes</b>.</p> <p>The FOAF profile</p>

<p>innovation that makes <b>the web more useful and more social.</b>" (Blogger Ben Parr)</p> <p>The protocol is <b>simple and minimalistic</b>, so is easily adaptable.</p> <p>An Open like system would be the ultimate goal of openness on the World Wide Web but, like so many open protocols before it, in order for it to work, big competitors like Microsoft, Google, Apple and Twitter will have to adapt it until then the <b>Open Graph by Facebook is the best available option.</b></p>	<p>profile data is full of duplicates and ambiguity.</p> <p><b>Lack of secondary attributes:</b> The Open Graph protocol only allows tagging of objects. While in comparison RDFa standard includes the ability to define relationships to objects. For example, a review of the book 'A' technically should not be tagged as an Open Graph object of the type book. In RDFa, the review could be tagged with the type "review" and then defined to have a relationship ("about") with the book 'A'. This will also cause a problem when two books with the same name would be considered to be the same book. A proper way to deal with this sort of thing is to introduce secondary attributes like 'writer' or a year that can help identify specific object, but the protocol does not define secondary attributes.</p> <p><b>Mark-up inside page.</b> There is no way to mark-up the objects inside the page. In its current version, the protocol only supports declaring that the entire page is about a person, an event, a book or a movie, but there is no way to identify objects inside the page. This is a use-case for bloggers and review sites. Each blog post typically mentions many entities and it would be nice to support this use-case from the start.</p> <p>The "open" graph is <b>not open</b>; you have to be a member of facebook to use it.</p>	<p>generated by Cheri is <b>Open.</b></p>
---	--	---



Despite the issues surrounding Facebook, bits of this platform bring together the visions of a social, personalised and semantic Web that has been discussed for a long time in the semantic Web communities and since del.icio.us pioneered Web 2.0 back in 2004.

User Content Observations (Text content mark-up and disambiguation):

Facebook's vision is both minimalistic and encompassing, yet this very minimalism and aim to encompass everything in and around Facebook (on Web), accompanied by the fast growing competition in the business, has resulted in the introduction of a premature and somewhat crude data graph with issues of data quality and cleaning to those who want to reuse and recycle it. The main issues found in the user generated data on Facebook gathered now through the open graph are listed in the right hand column of Table 4.2.1.1. However not all of these issues are unique to Facebook and are shared throughout most of the Web 2.0 platforms to some extent.

On its launch, neither Facebook nor the publishers (its partners) did any mark-up on their pages. At the time none of the entity pages on Facebook.com had Open Graph mark-up and thus Facebook's own pages remained closed. Ironically, this might not be because the company does not want to mark-up the pages, but it might be because it can't until it figures out what is actually on the page. This is what semantic technologies have been working on over the past several years. To be able to mark-up the pages correctly, especially the ones created by the users, Facebook needs to run them through a semantic processing and disambiguation process. In the current circumstances for our research the issues of disambiguation and semantic processing had to be performed by us to be able to power recommendations, to make social plugins and to facilitate good user experience. At the time of the submission of this thesis however Facebook is working hard to overcome these shortcomings and some mark-up is appearing on its pages and that of the partner companies such as IMDB and Pandora.

### 4.3 Finding the Right Vocabulary: A Universal Vocabulary or Specific vocabularies

A closer look at the user's social profile data reveals the heterogeneity of the content which leads to the problem of incorporating heterogeneous external communities of domain experts, for user profile data in order to resolve the issues such as those of disambiguation and proper semantic mark-up for the identified concepts as discussed in the previous section. To deal with this issue, we introduced the notion of a universal vocabulary. A universal vocabulary is a vocabulary that could act as the external communities of domain experts (a domain expert is a person with special knowledge or skills in a particular area of endeavour) for describing user profile data, in this way the aforementioned problem could be easily facilitated using the new representation. The use of a universal vocabulary to automatically and uniquely describe any resource, allows recommendations to be made even when the user has not previously expressed interest for any resources in a particular domain. The external resources that the users' data are to be compared to cannot be expected to expose information in the same format as that used to represent users within the system. Rather than attempting to obtain a mapping for each combination of users and domains, the notion of a universal vocabulary is introduced.

Of course, a vocabulary expressive enough to describe every conceivable resource cannot be expected to be readily available. Thus, the possibility of using DBpedia as an adequate replacement is explored in this section.

DBpedia is a community effort to extract structured information from Wikipedia and to make this information available on the Web. The DBpedia project extracts various kinds of structured information from Wikipedia editions in 97 languages and combines this information into a huge, cross-domain knowledge base (<http://blog.dbpedia.org/>).

Wikipedia was identified as a rich external source of information, due to its wide coverage of subjects, maintained by their respective interest groups (Giles, 2005). There has been much discussion on the need for the adoption of a shared language for the successful deployment of semantic technologies and the impossibility of imposing

a common, engineered understating of the world. DBpedia, as the largest dataset of structured information that is freely accessible and consensually built, can be regarded as a prominent candidate to fill this gap. This information may be deemed expert knowledge and the linked-data-graph spanned by Wikipedia articles together with the links between them can be used as the universal vocabulary. Each thing in the DBpedia data set is identified by a URI reference of the form <http://dbpedia.org/resource/Name>, where Name is taken from the URL of the source Wikipedia article, which has the form <http://en.wikipedia.org/wiki/Name>. Thus, each resource is tied directly to an English-language Wikipedia article. DBpedia provides three different classification schemata for things. A *Wikipedia Categories* (represented using the SKOS vocabulary and DCMI terms), the *YAGO Classification* (derived from the Wikipedia category system using Word Net) and *Word Net Synset Links* (generated by manually relating Wikipedia info-box templates and Word Net synsets and adding a corresponding link to each thing that uses a specific template. In theory, this classification should be more precise than the Wikipedia category system). Using these classifications within SPARQL queries enables to select things of a certain type.

In addition, the fact that articles are organised in categories provides added opportunities for extracting some semantics on the quality of the matching carried out. In order to project a resource onto Wikipedia we identify a page that corresponds to the resource. The projection then contains the aforementioned page along with pages that link to, or are linked from it. The projection contains the Wikipedia page that corresponds to the domain resource and any other pages connected to it via hyperlinks. Wikipedia does not actually contain a page for each resource in the world and using it as the Universal Vocabulary implies that some resources cannot be represented. Therefore the framework would be unable to predict the utility of such resources to the user. This shortcoming is inherent to DBpedia as well.

The DBpedia data set uses a large multi-domain ontology derived from Wikipedia. The data set currently describes 3.64 million “things” with over half a billion “facts” (as per July 2011) (<http://wiki.dbpedia.org/Datasets>). DBpedia uses the Resource Description Framework (RDF) as a flexible data model for representing extracted information and for publishing it on the Web. We use the SPARQL query language

to query this data. Following is an analysis of the Facebook user data with reference to DBpedia as a vocabulary. The aim of this preliminary analysis is to observe the feasibility of using DBpedia as a vocabulary for describing user generated content in Web 2.0. The details of the experiment are described as follows.

#### *4.3.1 Analysis of Facebook user data: Assessing the appropriateness of DBpedia as a universal vocabulary for Facebook user data.*

This experiment was run to analyse the appropriateness of the facebook user data as a user interest gathering resource and to evaluate the appropriateness of using DBpedia as a universal vocabulary for describing concepts in facebook user data. For this experiment we evaluated a data set from 186 Facebook user profiles. Each term in the data set was queried against DBpedia for successful hits. As mentioned earlier DBpedia uses the Resource Description Framework (RDF) as a data model for representing extracted information for the Wikipedia articles and for publishing it on the Web. We used the SPARQL query language to query each term in the facebook user profile against this data. DBpedia has a SPARQL query endpoint to write your queries against the DBpedia, but the process is very slow. Keeping this in mind a local copy of the DBpedia dataset was made on the local server and the SPARQL queries were made against it. The facebook user data comprised of the activities, interests, books, movies and music the users had showed their interest in on facebook. A SPARQL query was written against the DBpedia data set for each term in the Facebook user data (example of SPARQL query is given in section 5.2.3 under the heading ‘our approach’).

However it is seen that a single term in user’s facebook data may refer to multiple concepts in the DBpedia data set. This creates ambiguity that is the inability of pointing to the most appropriate concept amongst a set of related concepts. For example, ‘Paris’ may refer to many different concepts in the DBpedia (see [http://dbpedia.org/page/Paris\\_\(disambiguation\)](http://dbpedia.org/page/Paris_(disambiguation))). The ambiguity factor here in this experiment refers to the number of concepts a single term in facebook data points to when queried against DBpedia. On one hand this ambiguity prevents us from linking the facebook user term to the appropriate concept in DBpedia, but on the other hand it increases the chances of finding the appropriate concept showing us the diversity of

the DBpedia data set. The problem of ambiguous terms is answered in our thesis by first identifying the parent concept of the facebook user data term in question where ever possible and matching it with the parent term of the concept in DBpedia data set. If the information about the parent concept is not available with the user data then the user is presented with a list of the concepts in the disambiguation list of DBpedia and asked to choose the appropriate term. This method resolves the problem as can be seen from the evaluation in section 7.2.4.

The base of this evaluation was the principal of relevance, which states that; Something  $A$  is relevant to a task  $T$  if it increases the likelihood of accomplishing the goal  $G$ , which is implied by  $T$ . (Hjorland and Christensen, 2002). In our experiment  $A$  represents terms in the facebook user data,  $T$  is the task of querying each term against DBpedia concepts while  $G$  (goal) is getting the relevant concept URI. The data was queried against DBpedia as it is the largest and most diverse source of LD available on the Web.

Taking the principal of relevance one step farther we then calculated the term-ambiguity ratio for the each query performed. This gives us a true measure of the usability of the dataset. As mentioned earlier the user data is maintained in categories related to the context of origin of each term, to keep the context of use safe from being lost. Therefore for ease of understanding in Figure 4.3.1.1 the term-ambiguity statistics are cumulatively presented in their respective categories.

The term-ambiguity ratios were calculated using the percentage ambiguity function given below. Here ambiguity is defined as unclearness, by virtue of having more than one meaning.  $A_c$  is the percent ambiguity of data for a single category,  $c$  is the category for which the ambiguity is calculated,  $n$  is the total number of available terms in a category  $c$  with any number of successful hits in DBpedia and  $h$  is the total number of hits for a single term in this category.

$$A_c = \frac{100}{n} \sum_1^n \left( 1 - \frac{1}{h_i} \right)$$

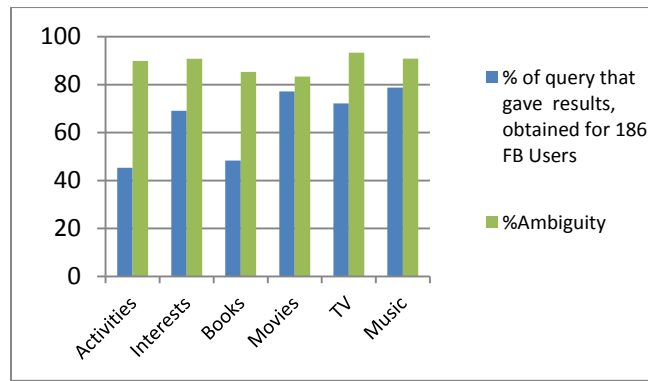


Figure 4.3.1.1: Successful DBpedia queries per category vs. % Ambiguity per category

Figure 4.3.1.1 describes the results from the queries done on DBpedia dataset to find relevant concepts for the Facebook user data. As mentioned earlier data from 186 Facebook users' profiles was gathered. The data gathered was already categorised into Activities, Interests, Books, Movies, TV and Music the user has shown interest in. Each term in these categories was individually queried upon DBpedia to find the related concept page in DBpedia (or in other words to find meaning of the term extracted from Facebook user data). The blue bar-chart in the above diagram represents the number of terms that resulted in successful queries, i.e., a related concept page was found against the query. The figure shows that about 69% of the queries in the 'interests' category were successful while almost 80% of the queries in 'movies' and 'music' categories were successful.

Apart from the successful queries, there were queries the DBpedia could not return a single concept page, but multiple related concept were returned. Such results were called ambiguous and are represented by green bars in the Figure 4.3.1.1. In this case a decision had to be made as to which concept was the correct concept representing the queried term.

It can be clearly seen from the ambiguity ratios that the social data need some work before it can be used for any deductive reasoning to generate intuitive results. The ambiguity ratios are quite high also because no pre-processing of data like spelling check, synonyms, spaces and multiple words or any other kind of data filtering technique is applied to the data. As the users social profile data is in essence free text data, it needs a certain amount of pre-processing to make it usable and less

ambiguous. However it is interesting to know that the majority of the concepts marked as ambiguous had the relevant concept listed as one of the suggestions in the ambiguity-list, i.e., the list of related concepts shown by DBpedia for that particular term e.g., the ambiguity list for the term ‘Paris’ can be seen at [http://dbpedia.org/page/Paris\\_\(disambiguation\)](http://dbpedia.org/page/Paris_(disambiguation)). This is a promising observation for our research as it indicates a strong potential in DBpedia to be used as a universal vocabulary to define users social data. However the direct use of this data at this stage does improve the quality of search to some extent as will be seen in section 4.4.1. This reinstates the fact that social data can be made searchable by referring to meaningful concepts in DBpedia.

## 4.4 The Cultural Heritage Domain Feasibility Analysis

Cultural Heritage online comprises of a plethora of diverse resources. Due to the diversity in the subject content of the CH items, their related history and the varied ways in which the information could be used, it is a very context intensive area. In order to provide meaningful personalization while searching CH content online interest terms in user profile need to be mapped to appropriate CH concepts, so that meaningful relations can be established between the user interest online and the CH concepts and personalised results could be suggested to the user. The purpose of this evaluation is to find the usability of Facebook user data to study if meaningful links between user interest terms and concepts in cultural heritage could be developed. So that meaningful recommendations can be made to the user when he/she is searching for something related to cultural heritage or when he/she visits an online museum The key evaluation goals are to assess social data for usability and findability of these terms against concepts in popular data sources and domain specific thesaurus which in this case are DBpedia (Auer, et al., 2007) and AAT (Paul, 2010) respectively. This will help with concept location, expansion and implicit reasoning during the query refinement and actual recommendation processes.

### 4.4.1 *Experiment Setup and Word/Data Analysis*

In order to establish a case for the use of social-data for personalization in the cultural heritage (CH) domain we conducted an experiment. There is a lot of structured and

unstructured information about cultural heritage on the Web and to extract something related, that might be of interest to the user is a cumbersome task. This is because, users are not conversant with the schemas used to store CH data nor are they aware of the concepts and properties modeling this context intensive domain. So it is hard for the user to formulate a query that might yield useful results. Our solution to the problem is to annotate user interest terms with the appropriate relevant concept and then use those relevant concepts for query refinement. To achieve this one must first know the *relevance* statistics of the user data to concepts in CH domain itself. To get an idea if the user interest could be used to suggest CH related data we ran a simple experiment comparing the user social profile interest terms to the concepts in a CH specific vocabulary, the Arts and Architecture Thesaurus (AAT). This will give us a rough estimate of the usability of social data for personalizing CH related search and recommendation activities. The results obtained from comparing user interest terms with AAT are then compared with the results of the same terms with DBpedia. The comparisons are shown in Figure 4.4.2.1. Note that ambiguity here refer to all the

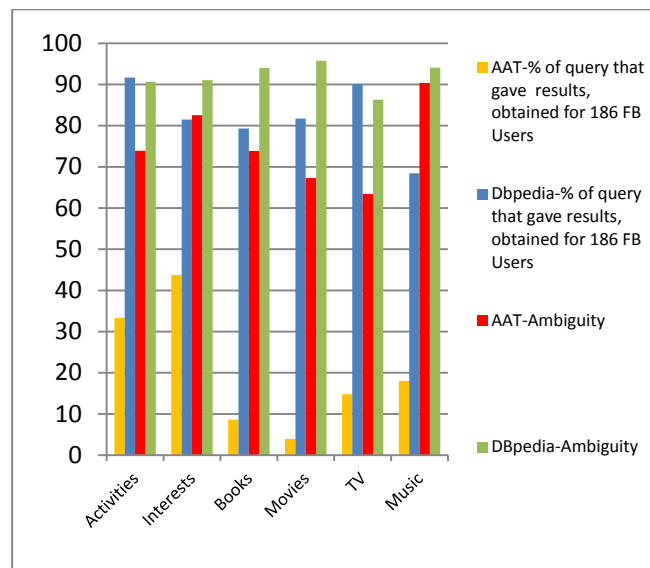


Figure 4.4.2.1: Successful AAT and DBpedia queries per category vs. % Ambiguity per category

cases where a term in user interest profile produced more than one relevant concepts in the AAT or DBpedia. This comparison was done on a newer version of the DBpedia hence the results obtained are significantly better than those obtained during the previous comparison with DBpedia shown in Figure 4.3.1.1. In this experiment we



queried the user data against AAT, as AAT is one of the most widely used structured vocabularies for arts and architecture concepts and compared it with the DBpedia query results from the previous experiment discussed in section 4.3.1. Figure 4.4.2.1 illustrates the graph that shows the comparison when a user interest term is searched in both Arts and Architecture Thesaurus and DBpedia. As evident from the bar graph the percentage of successful query results from AAT is far less than those of DBpedia but, if we look closely, the *interest* and *activities* categories are more comparable than the others, showing a potential mapping from general concepts of DBpedia to more CH specific concepts of the arts and architecture thesaurus (AAT). Another reason for using AAT in this evaluation is that, The Art & Architecture Thesaurus AAT, is a structured vocabulary that can be used to improve access to information about art, architecture, and material culture. The primary users of the AAT according to the Getty website include museums, art libraries, archives, and visual resource collection cataloguers, conservation specialists, and archaeological projects, bibliographic projects concerned with art, researchers in art and art history, and the information specialists who are dealing with the needs of these users. Thus finding a concept in AAT related to an interest term in a user profile insures that if a concept in AAT is comparable to a term in the user's interests, then a potential artefact could possibly be suggested to the user based on that concept. This will help in designing an interest refinement model that reasons upon and incorporates semantic networks that show links and paths between these concepts and others, since these relationships can make retrieval more successful. The graph also does the percentage term-ambiguity analysis for the AAT queries, using the same ambiguity formula as in the previous experiment developed earlier for DBpedia query analysis and compares the ambiguity ratios alongside the percentage hits.

These statistics show the complexity of identifying the correct concept to refer to, as a single term in a user profile may refer to multiple concepts of AAT and DBpedia thesaurus, thus making the term ambiguous. This leads us to the problem of identifying the actual concept the user was referring to while he/she wrote that term in his/her user profile, amongst a number of potentially right concepts that are identified in AAT and DBpedia. The context of the term, which is in this case the category in which the user has written the term in his/her profile (e.g. activities, books, movies, music, places etc.) help in narrowing down the correct concept to refer to. By

identifying whether the user is referring to Paris (Paris by Paris Hilton) the music album (if mentioned in the music category) or the movie Paris (if mentioned in movies/TV category) or the book Paris (if mentioned in the book category) or the place Paris. The categories of the terms in user profile are stored as metadata and can help in narrowing down to the correct concept.

### ***A Simple Search Experiment:***

SCENARIO: As an illustrative example we here design a scenario where one of the users (anonymous with an imaginary name Alan J. Smith) is on a business trip to London, he has a day before he flies back to states and wants to spend a few hours exploring the city for something he might be interested in. He does a simple query on Google (*Exhibitions London*). Table 4.4.1.1 gives the top eight results that he gets. We then refine the query with some of the interest terms that we gathered from her online profile and redo the query, the results of which are also mentioned in Table 4.4.1.1 This simple experiment of the use of Alan's profile to personalise her search, gives us a significant insight in the potential of social data on Web.

As the table shows, Query 1 returns reasonable results but leaves a lot of searching and clicking still for the user to do, before they can get something they might be interested in. In Query 2 the system incorporates some of the interests it attained from the users social Web profile (in this case facebook profile) and incorporates them to refine the original query thus helping the search engine in presenting the user with more personalised results and in doing so saving time and reducing chances of discontent and frustration.

Table 4.4.1.1: Search refinement using user's social Web data.

Query 1	Exhibitions London
Top 8 Outcomes	<ol style="list-style-type: none"> <li>1. Big <b>London exhibitions</b> in 2010 - Time Out <b>London</b></li> <li>2. Art galleries and <b>exhibitions</b> in <b>London</b> - Time Out <b>London</b></li> <li>3. <b>Exhibitions</b> in <b>London</b> in April 2010 - What's On - All</li> </ol>

	<p><b>In London</b></p> <ol style="list-style-type: none"> <li>News results for <b>exhibitions London</b> Top 10 Art <b>Exhibitions in London</b> - Visit <b>London</b></li> <li>Top 10 Art <b>Exhibitions in London</b> - Visit <b>London</b></li> <li><b>Exhibitions</b> and Trade Fairs - <b>exhibitions.co.uk</b></li> <li>Arts &amp; <b>Exhibitions</b>   <b>London</b> Evening Standard</li> <li>Home - EC&amp;O Venues   Event Venues in <b>London</b>   Conference centers ...</li> </ol>
<b>Query 2</b>	<b>Exhibitions London</b> (interest terms Applied)
Query refining term	(Dancing, Reggae, fashion, comedy)
Top 8 Outcomes	<ol style="list-style-type: none"> <li><b>Reggae</b> club nights in <b>London</b></li> <li>Shakespeare's Globe Theatre Tour and <b>Exhibition...</b></li> <li>Big <b>Dance</b> - Visit <b>London</b></li> <li><b>Reggae</b> Roast Warehouse <b>Dance</b> - 'Serious Single Launch Party ...</li> <li><b>London</b> Events   What's On In <b>London</b> Events Guide   <b>London</b></li> <li>Brent Black Music History Photographic <b>Exhibition</b></li> <li>Blaze street dance theatre show in London   <i>Raggs.com</i></li> <li>Irving Penn Portraits - Visit <b>London</b></li> </ol>

This simple experiment illustrates the potential that with meaningful contextual use and appropriate reasoning upon user's data this search can be farther improved to provide, more relevant search and recommendations and a better user experience.

## 4.5 Conclusions

This chapter presented a discussion to justify the use of Social Networks/ Web 2.0 data as a source of gathering user interest information in particular a discussion on Facebook as a Social Network of choice for this type of user data mining. After identifying the problems with the Web 2.0 user generated data such as the lack of proper mark-up, heterogeneity of data and disambiguation issues, this chapter proposes the idea of using appropriate vocabularies to resolve the identified issues.. DBpedia and AAT vocabularies were studied as potential vocabularies to describe concepts in user generated data.. This initial work provides a base for the choices we made to eventually create the idea of Cheri system (chapter 5) and to implement it as a recommender and search system (see chapter 6) according to the semantic Web and LD standards.

Section 4.2 discussed the possibilities of using social-media as a content mining facility for user interest gathering, for personalizing search and recommendation processes. Such information we believe can assist in reasoning and personalization activities in search and recommender systems especially in context intensive and task oriented scenarios.

Section 4.2.1 establishes a case for Facebook as a representative of a broad online population of individuals, whose online personal networks reflect their real world connections, making it an ideal environment to study user interest dynamics and information contagion, which is a useful phenomenon considering one of the aims of this research is to introduce the general Web users to cultural heritage related information according to their interests in a seamlessly unobtrusive yet pervasive manner. The discussion proceeds by giving an overview of the controversies and critiques surrounding facebook over the years. Table 4.2.1.1 highlights the history of Facebook's critiques, the issues different aspects of facebook brought forward and

amends made if any to overcome the issues. The most prominent among the issues were those of internal and external privacy threats, data ownership concerns and terms of use. This observation made us realise that as there is not as much of a strong privacy and data protection related policy inherent in Facebook as needed for our research; we need to take our own measures to ensure that the user data is asked for and used according to the user's permission in our system. Next a discussion on the reasons why facebook was chosen despite the inherent privacy concerns and critiques is summarised in Table 4.2.1.2. The popularity of Facebook, its status as the most used website over time and its strong growth in the last five years along with the strong statistics of traffic from Facebook to top newspaper sites which have nearly doubled since last year in Europe indicate the feasibility of Facebook as a suitable platform for distribution of knowledge. This makes it suitable for our research and indicates its potential for introducing cultural heritage related information to a vast number of Web users. In order to extract user information from Facebook we used the Facebook Connect API previously but later on with the introduction of the Facebook Open Graph protocol the data extraction model was redesigned to incorporate the changes. The Table 4.2.1.3 discussed the pros and cons of the new Facebook Open Graph protocol. The pros of the open graph include simplicity and minimalism and the possibility of synchronising user interest across the Web and making the Web more useful and social. Among the drawbacks the most prominent ones identified were the lack of support for object disambiguation and lack of proper mark up along with duplication of user data. To overcome these discrepancies the concept of a universal vocabulary was introduced in Section 4.3. A universal vocabulary is a vocabulary that can act as the external communities of domain experts for describing user profile data. In this way the aforementioned problems could be alleviated using this new representation.

DBpedia, as the largest dataset of structured information that is freely accessible and consensually built, can be regarded as a prominent candidate to fill this gap. This information may be deemed expert knowledge and the linked-data-graph spanned by Wikipedia articles together with the links between them can be used as the universal vocabulary. Next, in section 4.3.1, an experimental run on 186 Facebook user profiles in order to evaluate the feasibility of using DBpedia as the main vocabulary for

identifying user data concepts is discussed. The results of the experiment indicate, through the ambiguity ratios that the social data need some work before they can be used for any deductive reasoning to generate intuitive results. However the direct use of this data does improve the quality of the search to some extent as seen in section 4.4.1.

In figure 4.3.1.1 the ambiguity ratios are quite high because no pre-processing of data like spelling check, synonyms, spaces and multiple words or any other kind of data filtering technique is applied to the data. The user's social profile data is in essence free text data. It needs a certain amount of pre-processing to make it usable and less ambiguous. However it is interesting to know that the majority of the concepts marked as ambiguous had the relevant concept listed as one of the suggestions in the ambiguity-list. This is a promising observation for our research as it indicates a strong potential in DBpedia to be used as a universal vocabulary to define user's social data.

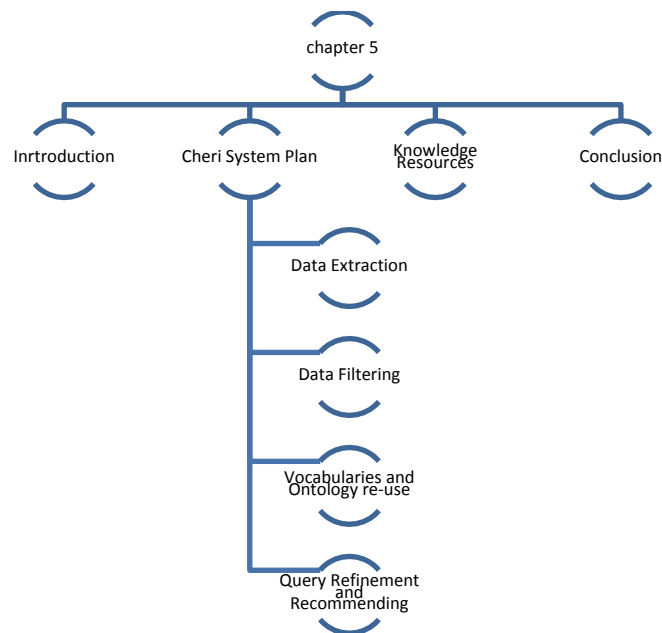
In order to establish a case for the use of social-data for personalization in the Cultural Heritage (CH) domain, Section 4.4 describes a simple social data analysis experiment and discusses its results. To get an idea if the user interest could be used to suggest CH related data the experiment compared the user social profile interest terms to the concepts in a CH specific vocabulary, i.e. The Arts and Architecture Thesaurus (AAT). AAT insures that if a concept in it is comparable to a term in the user's interests, than a potential artefact could be suggested to the user based on that concept. This gave us a rough estimate of the usability of social data for personalizing CH related search and recommendation activities. As evident from the bar graph the percentage of successful query results from AAT are far less than those of DBpedia but, if we look closely, the *interest* and *activities* categories are more comparable than the others, showing a potential mapping from general concepts of DBpedia to more CH specific concepts of the arts and architecture thesaurus (AAT).

The experiments conducted in Section 4.3.2 on social profile data from Facebook to establish a case for the use of social data and Section 4.4.2 to demonstrate the usability of this data for search refinement through personalization in the cultural heritage domain, provide the proof of feasibility needed to use users' social Web data from Facebook and the potential of DBpedia as an appropriate vocabulary for defining

this data. These decisions will eventually evolve into (by chapter 5) a cultural heritage oriented, term refinement and query optimization model which we believe will further improve the search and recommendation process by providing a suitable basis for improved query optimization.

## Chapter 5

# Modeling User-Interest Semantics



Chapter 5 topic hierarchy

This chapter describes our work in the semantic modeling of user interest information extracted from Web 2.0 sites. The requirements identified for adding semantics to user interest representations are presented. An introduction to the proposed framework design for a personalised recommender system that satisfies the requirements is introduced to be further discussed in chapter 6. The framework describes the process of annotating and enriching user interest information from Web 2.0 sites with LD information resources to produce an interest based user model characterised by an exportable and dynamic user interest profile. The integration of the user interest model thus obtained with the recommender system for providing personalised



recommendations is discussed. Finally the knowledge resources for provision of recommendations are identified.

## 5.1 Introduction to User-Interest Semantics

In this chapter we introduce the Cheri Project. Cheri is a project that provides user interest tailored information and recommendations about artwork from the V&A museum in London and recommendations related to the user's interest from LOD on Web. In Cheri, we used content from social Web profiles as context to recommend Cultural heritage related information as available through the V&A online collection and much more. This data is represented by ontologies and filtered using knowledge about user and user interest. Our aim is to improve cultural heritage information access and CH awareness among social Web users using the Cheri Project.

It is hard for the users to compose queries over complex domains like CH due to the vocabulary gap. Unfamiliarity of the content leads to ignorance to what could be found (information discovery), along with the enormous amount of information availability (information overload). The proposed model handles this by abstractly adding some domain specific concepts to the user query based on their interest profile (e.g. expanding the user query by adding the object types in the V&A that are most related to the user's query and interest) and search terms. In this way an abstraction over the original terms is provided; to help non-expert users in composing queries over complex and heterogeneous knowledge bases like cultural heritage while avoiding possible confusion.

The system works by initiating the generation of an extended-FOAF profile by seeding filtered concepts from the user's social network profiles and mapping them on to a universal vocabulary (DBpedia) thus giving meaning and a dereferenceable URI to the user interest terms (concepts). This extended FOAF profile, which is in essence an interest profile, is explained in section 5.2.3. The interest profile is made dynamic, that is it updates itself automatically as soon as the user indicates a new interest on their social network profile and logs into the Cheri system. It is also portable, as it complies with the standard FOAF ontology schema.

Cultural heritage is a field covering a wide range of content that varies significantly by type and properties, but is still semantically extremely richly interlinked. Presently, this content is still mainly in closed databases, distributed across national borders but with the advancements in web science more and more access points are becoming available online. The organizations managing these databases are of different kinds, such as museums, libraries, archives, media companies, and web 2.0 sites. Moreover, different natural languages and cataloguing practices are used in different countries and organizations. This creates the heterogeneity of CH resources which results in the need for a complex knowledge representation requiring an open, extensible data model.. In our research we use RDF as the common information model, and a collection of standard vocabularies and ontologies to enable semantic integration.

To overcome the technical and methodological barriers outlined earlier in chapter 4 we devised the following strategies. The implementation of these proposed strategies as discussed later on in this chapter and their testing in chapter 7 proved the usefulness of the approach. Following is a brief discussion on the challenges and the approach taken.

- ***Unobtrusive information gathering:***

A crucial part of our methodology was the provision of unobtrusive information gathering. This was ensured by gathering the publically available interest information about a user's online activities, which did not require any direct user involvement. This helped gather important information about the user's interest without requiring much help from the user, yet helped enrich the profile.

- ***Basic Concept Location:***

We used standard ontologies like FOAF and LOD resources like DBpedia, as shared vocabularies and thesauri to model the user's interest domain. This was achieved by linking interest terms used by users in their online social profiles after filtering them using different natural language processing techniques, to meaningful concepts in the above mentioned ontologies. This resulted in an ontology-based elicitation of user's interests and preferences, and was stored as an extended overlay context model or to be precise a user interest model.

- ***Removing the Vocabulary gap:***

By using standard ontologies like FOAF and DBpedia to conceptualise user interest terms we aimed to minimise the vocabulary gap. The concept-terms could then easily be mapped to more domain specific ontologies in order to support domain specific recommendations. We aimed to demonstrate this by mapping some of these concept-terms to a Conceptual Reference Model (CRM) which is a well-known and widely used ontology model in the cultural heritage domain, as will be discussed later. However this requires full access to the domain data or the knowledge resource as we are using the knowledge resources through APIs, which is a limitation of our system. A relatively simple domain ontology that represents the CH domain concepts in a simpler and more understandable form and fulfils the requirements of the current users is developed. This is implemented to represent the CH data related to the user interest and is named the Cheri Ontology.

## 5.2 Cheri System Plan

Following the approach suggested above and after an intensive literature review of the related technologies, the architecture for the *Cheri* system was proposed which is described in sections hereafter. The architecture proposed below intends to model user interest based on the social-Web profile owned by them. This model is then utilised in recommending cultural heritage resources that might be of interest for the user. Two systems are produced as a result; the Cheri recommender system and the Cheri search system, both of which are explained in chapter 6 in greater detail. Here we give an overview of the different modules that contribute towards the development of the two systems.

The proposed architecture consists of the following main components.

- ***Identifying a user's profile across a social network:***

The first module of the system identifies the user's social profile. This will help in deciding where to extract the user's data from.

- ***Data Extraction:***

This module describes a set of data extraction techniques mostly utilising public APIs provided by the data-owner sites themselves e.g., facebook, twitter, etc. and some open source scripts developed by individuals for the purpose of data extraction from different websites that don't have the facility otherwise.

- ***Data Filtering and Concept location:***

This module specifies a set of filters for cleaning the user data and making it usable for the next step.

- ***Concept mapping and Ontology re-use:***

This Module takes the set of filtered interest terms and equips them with semantics by categorization and use of well-known ontologies.

- ***User's interest-profile:***

This module describes a user interest profile and its underlying ontology, that helps relate the concepts from the user SNS profiles to the domain model to make recommendations.

- ***Recommender Domain module:***

This module describes how the user-interest model fits with the domain,by proposing ontology that helps relate the concepts from the user's interest profile to the domain objects.

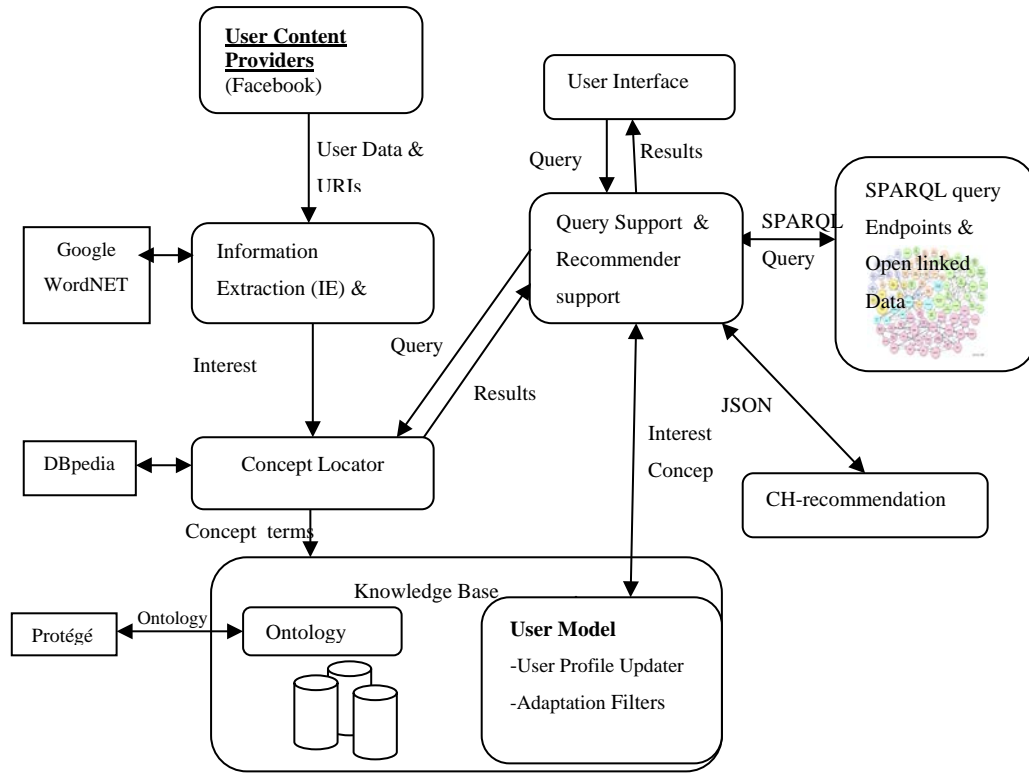


Figure 5.2.1: Cheri System Scheme

- **Query Refinement and Recommending module:**

The final portion is a query system and a recommender module that extracts data about the user's interest by querying our cultural heritage repository and LD on Web. In addition, this module takes as an input a set of 'interest concepts' and applies a concept expansion algorithm on them to discover new information for recommendations. The module also performs a query refinement function to discover the information most related to a user's interest.

A detailed discussion on the two Cheri systems is given in chapter 6. Here we present a description of various parts and functionalities of the architectural units.

### 5.2.1 Data Extraction Module

Users' information can be acquired in different ways, including: social tagging, query logging, and explicit user feedback. Users share a good proportion of personal data with proprietary databases in order to communicate with others in the network and

using the network itself creates a plethora of information. The internet traffic statistics state that, the total number of unique users visiting Facebook increased from about 90 million per month in April 2009 to 120 million by May 2010, and by 2012 the number further increased to 171.5 million, indicating the potential of user data contained in such websites. The data portability in social networks has been debated during the last few years. This data until recent years was mostly locked inside the network resulting in the loss of valuable information that otherwise could have assisted the users in exploring the Web. This information lock was once considered an advantage by the networks, but with the advent of social network technologies and their use this is generally not thought to be the case.

The Data Extraction module is generally responsible for collecting user related information, mostly user's interest information from the identified Social profiles or URIs. Most of the social networking sites have public APIs that provide mechanisms to enable the extraction of the user's public information, such as the <https://api.del.icio.us/v1/tags/get> in Del.icio.us, *flickr.tags.getListUser* method along with several others in Flickr, the *photos.getTags* method in Facebook along with many others, and the *user.getTopTags* method in Last.fm. Some of the sites like Flickr and Last.fm have nice public APIs that help retrieve a complete history of a user's tagging activities. However others are not as extraction friendly, so methods like screen scraping scripts need to be written. Thanks to the open source programming communities, scripts can be used off the shelf. Other projects such as (Szomszor, et al., 2008; Cantador, et al., 2008) have developed their own scripts for data extraction.

### **Our Approach:**

Our system previously used *Facebook Connect* to access publically available user data from Facebook that had a data storage restriction of no more than 24 hours. The site did not allow keeping a copy of the data extracted from it for more than a day. This gave little time for any complex data processing tasks to be performed on the data. Below in this section we give details of the Facebook Connect method of the data collection. However the system transferred to the Facebook graph after its launch and the details of its usage are discussed further in chapter 6. Though Facebook's open graph lifted the time restraints' it didn't go any further. A growing amount of user profile data remained full of duplicates and ambiguity. Facebook's Open graph

protocol did not resolve or support object disambiguation, or multiple objects on the same page, nor did it apply any mark-up on its pages that could be used directly at the time of launch. This called for additional data filtering, semantic processing and a disambiguation service as discussed in the following sections.

For the experiment our system collects the initial information about the user by connecting the user to their Facebook account at login time and extracting the publically available data fields from their Facebook profiles as they login. The information extracted from the profile is explicitly asked for at login time to keep the user aware of what information from their profile is being used. As explained in detail in chapter 6, in addition to this information the system also captures the current location of the user to provide the user with location based recommendations which are quite handy for example in a tourism scenario.

### ***Facebook Application:***

Initially a Facebook application was developed in order to facilitate the data extraction and introduce the social network users to the concept of the Cheri system. This mechanism allowed the user to add the application to their Facebook account so that the person's public data could be extracted. Later on the user was also able to use the application through his/her facebook account to navigate and use the Cheri recommender and search system.

The application developed used the Facebook Connect protocol to authorise the use of user data by directly asking for the user's permission. The details and screen shots of the process are given in Section 6.2.1. The `User.getInfo` method provided by Facebook was used to extract a variety of user specific information for each user, but the only storable values returned from this call were those related to user affiliations, `notes_count`, `proxied_email_address` and `profile_update_time`. Some basic information could also be extracted using `user.getStandardInfo`. We used `users_getinfo` method to extract information about a user as can be seen from the code below; following is some sample code used in the Facebook user data extraction method. Table 5.2.1 (a) gives the list of

information elements extracted from the user's facebook profile and their brief descriptions.

```
$fields=array('first_name','last_name','books','interests','current_location','education_history','has_added_app','movies','music');

$user_details=$facebook->api_client->users_getinfo($fbuid,$fields);
```

Table 5.2.1: (a) List of information elements extracted from a Facebook profile.

Info elements	Description
books	User-entered "Favourite Books" profile field. No guaranteed formatting.
interests	User-entered "Interests" profile field. No guaranteed formatting.
current_location	User-entered "Current Location" profile fields. Contains four children: city, state, country, and zip.
education_history	List of school information, such as education_info elements, each of which contains name, year, and child elements. If no school information is returned, this element will be blank.
has_added_app	[Deprecated] Boolean (0 or 1) indicating whether the user has authorised the application.
movies	User-entered "Favorite Movies" profile field. No guaranteed formatting.
Music	User-entered "Favourite Music" profile field. No guaranteed formatting.
profile_url	URL of the Facebook profile of the user. If the user has specified a username, the username is included in the URL, not profile.php?id=UID.
website	User-entered personal website profile field. No guaranteed formatting.

With Facebook Open Graph later on the following procedure and methods were used to extract the required fields of information.

### ***Facebook Graph:***

Facebook describes the Facebook Graph (Facebook Graph, 2012.) as the core of Facebook. The Graph API presents a view of the Facebook Social Graph, presenting the objects in the graph (e.g., people, events, photos, pages, etc.) and the connections between them (e.g., friendship relations, photo tags, etc.).



The objects of the Graph are accessed through unique IDs assigned internally by facebook. People and Pages with user names can be accessed through usernames as ID. All requests for access to the properties of the object are made using their unique identification number or username. Relationships between objects in facebook are called connections. The connections between the objects in facebook can be viewed using the URI structure ([https://graph.facebook.com/ID/CONNECTION\\_TYPE](https://graph.facebook.com/ID/CONNECTION_TYPE).) the set of connections the Facebook Graph holds between people and pages that we used in our system is listed in table 5.2.1 (b) below. Note the categories in facebook usually refer to a page on facebook describing the category.

Consider for example Salma N. Adil is a user of Facebook therefore there is a Facebook graph object that can be accessed through the Facebook Graph API using the URI structure (<https://graph.facebook.com/ID>). The API call to such a query will return the following structure:

```
{
  "id": "*****",
  "name": "Salma N. Adil",
  "first_name": "Salma",
  "middle_name": "N.",
  "last_name": "Adil",
  "gender": "female",
  "locale": "en_GB",
  "type": "user"
}
```

Now suppose we want to enquire what Salma's interests are. To explore this relationship through the Facebook Graph API we employ the URI structure ([https://graph.facebook.com/ID/CONNECTION\\_TYPE](https://graph.facebook.com/ID/CONNECTION_TYPE).) so the original query to the API will request the following URI ([https://graph.facebook.com/\\*\\*\\*\\*\\*/interests](https://graph.facebook.com/*****/interests)) and the results will be as follows.

```
{
  "data": [
    {
      "name": "Flowers",
      "category": "Interest",
      "id": "114937881856580",
      "created_time": "2011-07-04T22:41:55+0000"
    },
    {
      "name": "Horses",
      "category": "Animal",
      "id": "111933198826503",
      "created_time": "2011-07-04T22:36:10+0000"
    },
    {
      "name": "Puzzles",
      "category": "Interest",
      "id": "108089669223594",
      "created_time": "2011-07-04T22:25:52+0000"
    },
    {
      "name": "Cars",
      "category": "Interest",
      "id": "110962938928704",
      "created_time": "2011-07-04T22:25:48+0000"
    }
  ]
}
```

We can see here that with the use of the Facebook Graph API the extraction of user interest has been made much simpler than before, using the LD concepts. To take things further we can see that Salma is interested in flowers. If we query and see what the object “flower” refers to in the facebook Graph we would query the Facebook Graph API for the following URI (<https://graph.facebook.com/114937881856580>). The results would be shown as follows. We can see from the results that a concept in Facebook Graph in essence is linked to a facebook page describing that object. This follows the LD principle of representing each object with a URI, but it does not make it open; as the URI is an internal URI to Facebook. Our work in the presence of Facebook Graph still holds unique as it links user interests extracted from the Facebook Graph API to an open universal vocabulary, DBpedia.

```
{
  "id": "114937881856580",
  "name": "Flowers",
  "link": "http://www.facebook.com/pages/Flowers/114937881856580",
  "likes": 195182,
  "category": "Interest",
  "is_published": true,
  "is_community_page": true,
  "description": "<p>A <b>flower</b>, sometimes known as a bloom or <a href=\"/pages/w/10",
  "talking_about_count": 521,
  "type": "page"
}
```

The dereferenceable links made with Wikipedia pages to describe the interest through DBpedia are an open resource to the Web rather than a closed community or network.

Table 5.2.1: (b) List of information elements (connections) extracted from a Facebook user (object) profile by our system.

<b>Name</b>	<b>Description</b>	<b>Return</b>	<b>Description</b>
interests	The interests listed on the user's profile.	user_interests or friends_interests.	array of objects containing interest id, name, category and create_time fields.
location	The user's current city	user_location or friends_location	object containing name and id
likes	All the pages this user has liked.	user_likes or friends_likes.	array of objects containing like id, name, category and create_time fields.
movies	The movies listed on the user's profile.	user_likes or friends_likes.	array of objects containing movie id, name, category and create_time fields.
books	The books listed on the user's profile.	user_likes or friends_likes.	array of objects containing book id, name, category and create_time fields.
music	The music listed on the user's profile.	user_likes or friends_likes.	array of objects containing music id, name, category and create_time fields.

The Graph API allows access to all the public information about an object (e.g., user). For example, <https://graph.facebook.com/ID> (Salma N. Adil) returns all the public information about Salma, i.e., a user's first name, last name and gender are publicly available as can be seen from the above example. However to get additional information about a user, you must first get their permission. At a high level, you need to get an *access token* for the facebook user. After you obtain the access token for the user, you can perform authorised requests on behalf of that user by including the access token in your Graph API requests: in the following format ([https://graph.facebook.com/220439?access\\_token=](https://graph.facebook.com/220439?access_token=))

For example ([\(https://graph.facebook.com/\\*user id\\*?access\\_token=...\\_](https://graph.facebook.com/*user id*?access_token=..._) (Salma N. Adil) returns additional information about Salma as can be seen in the following API response.

```
{
  "id": "448822222",
  "name": "Salma N. Adil",
  "first_name": "Salma",
  "middle_name": "N.",
  "last_name": "Adil",
  "link": "http://www.facebook.com/profile.php?id=448822222",
  "location": {
    "id": "108426499181164",
    "name": "Southampton"
  },
  "quotes": "For my part, I know nothing with any certainty, but the sight of the stars",
  "education": [
    {
      "school": {
        "id": "115597998461181",
        "name": "burnhall"
      },
      "year": {
        "id": "143018465715205",
        "name": "2000"
      },
      "type": "High School"
    },
    {
      "school": {
        "id": "112164985476992",
        "name": "Southampton University"
      },
      "year": {

```

### 5.2.2 Data Filtering and Concept Location

The data from the social Web comes with some inherent problems, limitations and weaknesses that need to be addressed in order to make it useable. Some of the major issues here are those of ambiguity, spaces, and multiple words, synonyms (Mathes, 2004.) and typographical errors or misspellings. Most of the techniques employed in solving these problems have their origin in Natural Language Processing and Information Retrieval.

#### Spelling Correction

Users make spelling errors either by accident, or because the concept they are expressing for has no definite spelling to their knowledge. In practice and in the literature, normally a spelling corrector utilises one of several methodologies and steps to provide a spelling suggestion (Eulerfx, 2009). Some of the more common ones are listed as below:

- The first step is to deduce a way to identify whether spelling correction is required. These may include insufficient results, or results which are not specific or accurate enough according to some measure.
- Next a large and authentic resource of text or a dictionary, where all, or most of the words are known to be correctly spelled, is employed to identify the best suggestion word which is the closest match based on one of several measures. The most intuitive method to do this is by identifying similar characters. In research and practice two (bigram) or three (trigram) character sequence matches are found to work best. To further improve results, a higher weight is applied for a match at the beginning, or end of the word. To improve the performance, all the words are indexed as trigrams or bigrams, when a lookup is performed, the system adopts the n-gram technique, and lookup via hash-table or trie (prefix tree) is performed (Pauls and Klein, 2011).
- Use of heuristics related to potential keyboard mistakes based on character location is a much used technique. For example "hwlllo" should be "hello" because 'w' is close to 'e'.
- Phonetic keys such as Soundex or Metaphone are sometimes used to index the words and look up possible corrections. In practice this normally returns worse results than using n-gram indexing.
- Whatever combination of methods and techniques may be applied in the end the decision is to select the best correction from a list. This may be done by use of a distance metric such as Levenshtein (1966), the keyboard metric, etc.
- For a multi-word phrase, only one word is misspelled, in which case the remaining words are used as context in determining a best match.

## **Our Approach**

Google being the largest Web search engine has excellent algorithms for statistical language processing problems such as spelling correction to help improve the search results. Our system develops a mechanism to employ Google's "did you mean" spelling correction mechanism to locate the right concepts in the DBpedia vocabulary linking it to the corresponding Wikipedia page URI.

### Google Spelling Correction:

Google *did you mean* PHP Class is a compact but very powerful class when it comes to spelling checking and suggestions. It uses the *did you mean* feature to return filtered data. The underlying method involves a thorough check of the common occurrences of a term to see if the most common version of the word's spelling is used. If the analysis shows that a particular version of the word is likely to generate more relevant search results with a certain spelling, it will ask 'did you mean' while suggesting the more common spelling. Because Google's spellcheck is based on occurrence of all words on the internet, and because it constantly checks for possible misspellings and their likely spellings by using words it finds while searching the Web and processing user queries, it is able to suggest common spellings for proper nouns (names and places) and other terms that might not appear in a standard spelling check program or a dictionary. This leads us to the conclusion that though Levenshtein, Soundex or the LIKE function (The LIKE function determines if a character expression matches another character expression. In SQL the LIKE operator is used in a WHERE clause to search for a specified pattern in a column) have their own significance they cannot substitute for the machine learned data from humans as in the case of Google. This analysis makes the Google 'did you mean' the best candidate algorithm to be used in our system.

### ***Other Google Tools:***

Other interesting tools that Google has are Google Search API and Google Custom Search API. The Google Web Search API lets you put Google Search in your Web pages with JavaScript. You can embed a simple, dynamic search box and display search results in your own Web pages or use the results in innovative, programmatic ways. Google Web Search API has been officially deprecated as of November 1, 2010. Now Google has shifted to The Google Custom Search API which lets you develop websites and programs to retrieve and display search results from Google Custom Search programmatically. With this API, you can use RESTful requests to get either Web search or image search results in JSON or Atom format.

## **Lexical Analysis**

An interest term extracted from the SNS might be expressed in different word forms, or plural and singular may exist. The type of method we use for concept location using Google caters for the plural and singular terms. In practice the data might also need some processing to convert any special characters that may occur to a base form (e.g., ö to o). Most of the user data extracted from SNS for this project is extracted from fields designed to deal with single words or a few words only, e.g., user interests are usually expressed as single or multi words. This is the case for most social tagging systems. But often service restricts the use of multiple words especially in tagging system, e.g., this problem was observed in del.icio.us. As the site doesn't allowed space in a tag in order to restrict user from using multiple words to tag an object. But it was observed that the users still entered multiple words written as a single string without spaces, such strings are hard to comprehend. However Facebook allows multiple word values and spacing.

Common stop-words such as pronouns, articles, prepositions, and conjunctions are removed in general as the final step in data filtering. In our case as the data is in the form of single or multi word phrases and very short sentences, the removal of stop words is not required, as it may prohibit the location of the actual concept e.g. in movie/book names this could result in a wrong concept location.

Some of the related work in data filtering is done in the keyword extraction research from documents (Frank, 1999; Turney, 2000) and Web pages (Yih, Goodman, Carvalho, 2006). However, it has been shown statistically that the social network data snippets are extremely short and more informal (Li, et al., 2010). Accordingly insights from related work on traditional data filtering do not hold true in such cases as (Li, et al., 2010) and the work presented in this theses.

## **Concept location**

Using our Link Generation Service, for each interest identified, we use a technique to identify the respective article on Wikipedia describing the interest concept. This approach takes as an input the corrected term obtained through the Google 'do you mean' algorithm. The URI of the page describing the term-concept is achieved by performing a Google search and restricting it to the Wikipedia domain. This is

accomplished by adding `site:en.wikipedia.org` to the query string. The methodology is further explained by example in chapter 6.

A similar approach Using the Yahoo REST API is followed by (Biddulph, 2005) and (Sinclair, Lewis and Martinez, 2007). Their approach returns the first result of the Google query, which will typically be the Wikipedia article describing the person. However, there might be cases when this will not happen. Our approach is a modification of their approach. From observation we have realised that if the wiki link is not the first link in the search results it is one of the top 10 results. So to improve upon the probability of finding the right link, we search the top ten results thus obtained for the wiki term and pick the first result having Wikipedia mentioned in its URI. Our approach has the following steps;

- Find the correct spelling/form of the extracted user interest term.
- Google Search the term using a restricted query structure achieved through concatenation of the (`en.wikipedia.org`) with the search string. Figure 6.2.2.2 describes the procedure in detail.
- Scan the top ten results for the term ‘wiki’ to insure that the linked page is always a wiki link.
- Create the link.

The procedure is quite simple but the result is a surprisingly powerful dynamic link generation service that, through a Web 2.0-style mash-up approach, uses the rich content from Wikipedia as its underlying link-base. Unlike traditional link services, where the links in the link-base are typically defined in advance, the system is able to dynamically add links to any concept described on Wikipedia.

### 5.2.3 *Vocabularies and Ontology Re-Use*

Once the data is filtered and cleaned to make it semantically sound and formally represented in machine readable format we employ the use of the LD standards as stated below: (Berners-Lee, 2006)

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 RDF information.
4. Include RDF statements that link to other URIs so that they can discover related things.

We use standard ontologies and classifications to describe the user data and their relationships thus ensuring data portability. This process will eventually lead to a dynamic, portable, machine-readable user interest profile, which we will use in our system for providing personalised recommendations. The development model is discussed as follows:

### **Ontology Re-Use**

One rarely has to start from scratch while developing ontologies. There is almost always an ontology available from a third party that can help provide a useful start. Thanks to the efforts of last few decades, there is now a wide variety of classifications, vocabularies, taxonomies and ontologies available to choose from; for example *coded bodies of expert knowledge*, which in the case of cultural heritage are for example the Art and Architecture Thesaurus (AAT)<sup>1</sup> containing around 34,000 concepts, the Union List of Artist Names (ULAN)<sup>2</sup> enlisting 127,000 record entries on artists, the Iconclass vocabulary of 28,000 terms for describing cultural images<sup>3</sup>, and in the geographical domain the Getty Thesaurus of Geographic Names (TGN)<sup>4</sup>, containing around 1,115,000 records. While, most of the ontological efforts have also been made towards the more domain specific ontologies, attempts have been made to define generally applicable ontologies, sometimes known as upper-level ontologies. Some examples include Cyc<sup>5</sup>, Standard Upper level Ontology SUO<sup>6</sup>, Yago<sup>7</sup> and the DBpedia ontology. There are also resources that are simply sets of terms loosely

---

<sup>1</sup> [http://www.getty.edu/research/conducting\\_research/vocabularies/aat/](http://www.getty.edu/research/conducting_research/vocabularies/aat/)

<sup>2</sup> [http://www.getty.edu/research/conducting\\_research/vocabularies/ulan/](http://www.getty.edu/research/conducting_research/vocabularies/ulan/)

<sup>3</sup> <http://www.iconclass.nl/>

<sup>4</sup> [http://www.getty.edu/research/conducting\\_research/vocabularies/tgn](http://www.getty.edu/research/conducting_research/vocabularies/tgn)

<sup>5</sup> <http://www.opencyc.org/>

<sup>6</sup> <http://suo.ieee.org/>

<sup>7</sup> <http://dmoz.org/>

organised in a specialised hierarchy often known as *Topic Hierarchies*. These are not strictly taxonomies but rather a mixture of different specialization relations like *is-a*, *contained in*, *part of*, relations. Examples of such resources include the Open Directory Hierarchy<sup>7</sup> which contains more than 400,000 categories and is available in RDF format. Similarly some *linguistic resources* like WordNet<sup>8</sup> are successfully used as a nice starting resource for ontology development.

### **Using DBpedia for an Uncontrolled Vocabulary and term Ambiguity:**

Social network systems are mostly based on uncontrolled vocabularies i.e., there are usually no guidelines or scope definitions or precedence, to assist the user. Therefore, users may refer to different resources with the same word meaning different things at different places and vice versa. Similarly users might use acronyms that can be expanded in different ways. As SNS data are in most cases simple words with no semantics or hierarchical structure, this might result in a set of unresolved concepts. These issues together result in the loss of potentially useful data and therefore should be dealt with as much as possible.

Users should be given the capability to execute complex queries in order to provide greater accuracy in their searching endeavours over the Web. The semantic Web promises to provide such a feature by making the concepts within data explicit. To start with as discussed in chapter 4 we have chosen DBpedia as a universal vocabulary for user content classification and disambiguation. The feasibility study and the initial experiments with DBpedia were done in support of this decision (as seen in chapter 4).

Apart from disambiguating and adding semantics to the user data, DBpedia is also used as an LOD resource to expose and recommend interest related information to the user.

We also developed a number of techniques to improve the accuracy of the system using the DBpedia ontology. For example for automatic disambiguation one of the techniques employed is that the system checks that the page identified as the resource page by our automatic link generator is actually an article about an interest by looking

at the genera category of the DBpedia ontology on the page, and matching it with the category (if mentioned) from which the interest was originally extracted from the SNS (facebook) and matching them.

The automatic predicating technique described in chapter 6 is also based on the DBpedia ontology. To start with we have focused on extracting and filtering user interests from SNS, and our aim is to then retrieve structured information from the Wikipedia article (i.e., related DBpedia entries) to augment our knowledge base.

### ***Our Approach:***

The filtered interest terms from the user's online social network profile are mapped to concept in the DBpedia ontology; however the category list from WordNet was used with Wikipedia wherever possible as it is more structured than Wikipedia in its hierarchy. We suggest that cleaned-up user social data when categorised and mapped to standard ontologies following the LD principles can express the user interest more accurately and in a domain independent and reusable manner.

Before we proceed farther, we will give an example, to illustrate how the terms extracted from the user SNS profile can be mapped to the DBpedia property values and how they are expanded to acquire useful information entries for the *user profile*.

Consider Bia has the book “The Lord of the Rings” mentioned as her favourite book in her Facebook profile.

<books>

Alice's Adventures in Wonderland, **The Lord of the Rings**, Through The Looking Glass, The picture of dorian gray, La tahzan (Arabic), Kalila wa dimna (كليلة و دمنة - Arabic), Le petit prince (French), Les Femmes savantes(French)

</books>

After identifying the concept URI for the given user interest, by the above mentioned concept location mechanism, we SPARQL query the DBpedia for the book The Lord of The Rings, through the SPARQL endpoint (<http://dbpedia.org/snorql/>).

PREFIX dbpedia2: <http://dbpedia.org/property/>

```
SELECT * WHERE {  
  ?obj dbpedia2:name ?name.  
  ?obj dbpedia2:genre ?gen.  
  ?obj dbpedia2:author ?auth.  
  ?obj dbpedia2:id ?id.  
  ?obj dbpedia2:mediaType ?mediatype.  
  ?obj dbpedia2:precededBy ?pBy.  
  ?obj dbpedia2:books ?bk  
  FILTER (regex (?name,"The Lord of the Rings","i"))  
}
```

Following are some values from the DBpedia entry which we get as a result.

<u>dbpprop:author</u>	dbpedia:J._R._R._Tolkien
<u>dbpprop:books</u>	dbpedia:The_Two_Towers dbpedia:The_Fellowship_of_the_Ring dbpedia:The_Return_of_the_King
<u>dbpprop:country</u>	dbpedia:Literature_of_the_United_Kingdom
<u>dbpprop:genre</u>	dbpedia:Adventure_novel dbpedia:High_fantasy dbpedia:Heroic_romance
<u>dbpprop:id</u>	46316 (xsd:integer)
<u>dbpprop:mediaType</u>	Print
<u>dbpprop:name</u>	The Lord of the Rings
<u>dbpprop:precededBy</u>	dbpedia:The_Hobbit

By the careful selection of the right properties to add to the user model an intuitive and useful set of values can be obtained which can provide a useful insight into what the user may or may not like. For example; in this case `dbpprop:genre` can help identify that our user may be interested in Adventure novels, and fantasy, when it comes to book reading. The `dbpprop:genre` for all his books could be checked and weight to find which sort of books/topics he/she is most interested in. Similarly fields like `dbpprop:books` and `dbpprop:preceded by` can be used to make related book recommendations to the user.

### Using FOAF to represent user data

In the current Web 2.0 landscape most of the social networking sites and services do not facilitate connection amongst users of other similar services. This is also the case with user data portability and profiling. This is mainly because every service has its

own data representation that is invisible to other systems that provide social networking facilities. Thus different users remain enclosed in different networks forming disconnected components of the global social network as if living on isolated islands (Frivolt and Bielikova, 2006). Even the same users across different social networks have scattered or duplicated identities. This is mainly because profiles stored inside current social networking systems are not addressable, whereas RDF has its Uniform Resource Identifiers and those can be reused by any service that has access to them. Keeping this in mind we use an RDF based format FOAF for the user profile description. Table 5.2.3.1 shows a strong possibility that FOAF will become a standard for providing personal information on the Web. However there is a need for the existing personal profiles residing in different social networking sites to be brought to the common standard and vocabularies for a representation such as FOAF. In addition the sensitive parts of the FOAF profiles should be ensured against non-authorised viewing. Although the later issue is out of the scope of our research, we do discuss it in this section. However we will focus more on the user's interest data extracted from the social networking sites being represented in a standard and reusable format like FOAF and RDF.

We will start our discussion with a few concerns with the FOAF ontology and proceed from there to a discussion on the significance of FOAF to its creators and consumers. FOAF has been criticised for not being able to deal with user content privacy, partially the lack of privacy in the social network services is mirrored through in the FOAF profiles. Concerns have been raised regarding the FOAF profiles to be more prone to spamming than the social networking sites themselves. Nasirifard, Hausenblas and Decker (2009) illustrate by example fake attacks using information from users' FOAF profiles. They argue that crawling heterogeneous and highly customised social networking sites for finding users' email offers a huge overhead for the spammers. In addition users may create fake user profiles with incomplete or pseudo names on the social networks that may be of not much use to the spammers. FOAF, on the other hand, is hosted on personal Web pages and is generated automatically from reliable user data. There have also been criticisms about FOAF unique ids encouraging incidental unauthorised record merging (<http://wiki.foaf-project.org/w/Criticism>), However as FOAF is kept in an open extendable format, with time, appropriate

solutions and ways to bypass these shortcomings were suggested, such as by (Frivolt, and Bielikova, 2007).

In our case FOAF was chosen to represent user interest information gathered from social networking sites keeping in mind the significance of the format for the information creators and the information consumers which can be illustrated by the work of Dumbill (2002a; 2003) and is summarised in the following table.

Table 5.2.3.1: Reasons for choosing FOAF

Usefulness of FOAF	Reasons
For Information creators,	FOAF helps in <b>Managing communities</b> by offering a basic expression for community membership. Many communities have flourished on the Web, e.g., companies, professional organizations, social groups.
	Helps in <b>Expressing identity</b> by allowing unique user IDs across applications and services <b>without compromising privacy</b> . For example, the <i>foaf:mailbox_sha1sum</i> property contains the ASCII-encoded SHA1 hash of a mailbox URI. The encoding is designed as a one-way mapping and cannot be trivially reverse-engineered to reveal the original email address. Thus prevent others from faking an identity.
	<b>Indicating authorship.</b> FOAF tools use digital signatures to link an email address with a document. Commonly, Open-PGP is used, along with the new namespace <i>http://xmlns.com/wot/0.1/</i> to denote concepts. Thus forming a “Web of trust”. This process associates a signature with the document itself and then specifies a signature for the linked document as part of a <i>rdfs:seeAlso</i> link. In this way, authorship information can be expressed both inside and outside of the concerned documents.
FOAF supports consumers by:	Allowing <b>provenance tracking</b> and accountability; On the Web, the source of information is just as important as the information itself in judging its credibility. Provenance tracking RDF tools can tell where and when a piece of information was obtained. A practice common to the FOAF community is to attach the source URI to each RDF statement.
	<b>Providing assistance to new entrants</b> in a community. For example, people unfamiliar with a community can learn the structure and authority of a research area from the community’s FOAF files.
	<b>Locating people with common interests.</b> Users tend to have interests and values similar to those they desire in others (Adamic, Buyukkokten and Adar, 2003). Peer-to-peer relationships are an essential ingredient to collaboration, which is the driving force of online communities.
	<b>Augmenting email filtering</b> by prioritizing mail from trustable colleagues. Using the degree of trust derived from FOAF files, people can prioritise incoming email and thus filter out those with low trust values.
Some other	Among a large number of ontologies that have been published on the Web, only a few are well populated, i.e., have been brought to any significant use.

useful statistics	<p>An investigation of the namespaces of well-populated ontologies (see Table 5.2.4.3) by Frivolt and Bielikova (2007) revealed that, besides the meta-level ontologies (i.e. RDF, RDFS, DAML and OWL), one of the best populated ontologies is FOAF (Friend-of-a-Friend).</p> <p>Table 5.2.4.3 Eight best Populated ontologies</p> <p><b>EIGHT BEST POPULATED ONTOLOGIES (GENERATED IN JUNE,2004)</b></p> <table><tr><th>Onto. Name</th><th>Namespace URI</th><th># of Docs. Populated</th></tr><tr><td>RDF</td><td><a href="http://www.w3.org/1999/02/22-rdf-syntax-ns#">http://www.w3.org/1999/02/22-rdf-syntax-ns#</a></td><td>&gt; 1,129,749</td></tr><tr><td>FOAF</td><td><a href="http://www.foaf-project.org/">http://www.foaf-project.org/</a></td><td>&gt; 1,126,002</td></tr><tr><td>DC</td><td><a href="http://purl.org/dc/elements/1.1/">http://purl.org/dc/elements/1.1/</a></td><td>&gt; 1,117,433</td></tr><tr><td>RDFS</td><td><a href="http://www.w3.org/2000/01/rdf-schema#">http://www.w3.org/2000/01/rdf-schema#</a></td><td>&gt; 1,129,749</td></tr><tr><td>MCVB</td><td><a href="http://webns.net/mvcb/">http://webns.net/mvcb/</a></td><td>&gt; 8,838</td></tr><tr><td>RSS</td><td><a href="http://purl.org/rss/1.0/">http://purl.org/rss/1.0/</a></td><td>&gt; 7,560</td></tr><tr><td>vCard</td><td><a href="http://www.w3.org/2001/vcard-rdf/3.0#">http://www.w3.org/2001/vcard-rdf/3.0#</a></td><td>&gt; 6,229</td></tr><tr><td>Bio</td><td><a href="http://purl.org/vocab/bio/0.1/">http://purl.org/vocab/bio/0.1/</a></td><td>&gt; 6,183</td></tr></table>	Onto. Name	Namespace URI	# of Docs. Populated	RDF	<a href="http://www.w3.org/1999/02/22-rdf-syntax-ns#">http://www.w3.org/1999/02/22-rdf-syntax-ns#</a>	> 1,129,749	FOAF	<a href="http://www.foaf-project.org/">http://www.foaf-project.org/</a>	> 1,126,002	DC	<a href="http://purl.org/dc/elements/1.1/">http://purl.org/dc/elements/1.1/</a>	> 1,117,433	RDFS	<a href="http://www.w3.org/2000/01/rdf-schema#">http://www.w3.org/2000/01/rdf-schema#</a>	> 1,129,749	MCVB	<a href="http://webns.net/mvcb/">http://webns.net/mvcb/</a>	> 8,838	RSS	<a href="http://purl.org/rss/1.0/">http://purl.org/rss/1.0/</a>	> 7,560	vCard	<a href="http://www.w3.org/2001/vcard-rdf/3.0#">http://www.w3.org/2001/vcard-rdf/3.0#</a>	> 6,229	Bio	<a href="http://purl.org/vocab/bio/0.1/">http://purl.org/vocab/bio/0.1/</a>	> 6,183
Onto. Name	Namespace URI	# of Docs. Populated																										
RDF	<a href="http://www.w3.org/1999/02/22-rdf-syntax-ns#">http://www.w3.org/1999/02/22-rdf-syntax-ns#</a>	> 1,129,749																										
FOAF	<a href="http://www.foaf-project.org/">http://www.foaf-project.org/</a>	> 1,126,002																										
DC	<a href="http://purl.org/dc/elements/1.1/">http://purl.org/dc/elements/1.1/</a>	> 1,117,433																										
RDFS	<a href="http://www.w3.org/2000/01/rdf-schema#">http://www.w3.org/2000/01/rdf-schema#</a>	> 1,129,749																										
MCVB	<a href="http://webns.net/mvcb/">http://webns.net/mvcb/</a>	> 8,838																										
RSS	<a href="http://purl.org/rss/1.0/">http://purl.org/rss/1.0/</a>	> 7,560																										
vCard	<a href="http://www.w3.org/2001/vcard-rdf/3.0#">http://www.w3.org/2001/vcard-rdf/3.0#</a>	> 6,229																										
Bio	<a href="http://purl.org/vocab/bio/0.1/">http://purl.org/vocab/bio/0.1/</a>	> 6,183																										
	<p>Finally, representing personal information is a popular theme in ontology engineering. As reported by Swoogle* more than 1,000 RDF documents have defined terms containing ‘person’. The other well populated non-meta ontologies in Table 5.2.4.3 include: DC (Dublin Core Element Set), which defines document metadata properties without domain/range qualification, and RSS** (RDF Site Summary), which is “a lightweight multipurpose extensible metadata description and syndication format” for annotating websites.</p> <p>*This is reported by Swoogle (<a href="http://swoogle.umbc.edu">http://swoogle.umbc.edu</a>), a RDF crawling and indexing engine.</p> <p>** Swoogle had discovered approximate 80,000 RSS documents by September, 2004.</p>																											

We conclude from the above discussion that FOAF (Friend-Of-A-Friend) is an ontology for expressing information about persons and their relationships. FOAF takes the social networking aspect of the Web future by enriching the expression of personal information and relationships. So it is a useful building block for creating information systems that support online communities (Dumbill, 2002b). It is a collaborative project that has evolved into a flexible and practically used ontology. The FOAF vocabulary is identified by the namespace URI <http://xmlns.com/foaf/0.1/>. It defines both classes (e.g., foaf: Agent, foaf: Person, and foaf: Document) and properties (e.g., foaf: name, foaf: knows, foaf: interests, and foaf: mbox) in RDF format. In contrast to a fixed standard, the FOAF vocabulary is managed in an open source manner, i.e., it is open for extension. Therefore, inconsistent FOAF vocabulary usage is expected across different FOAF documents. We believe that person's FOAF

profile can be extended with additional information to express a person's interest in cultural heritage.

***Our extended ontology:***

In order to provide personalization we need to take into account the users of the system, their interests and attributes related to the interest. To accommodate this aspect of the system requirement, we extend the foaf: interest property of the FOAF ontology with additional concepts. This extension is referred to in our system as CH\_INTEREST vocabulary, which is used to model the use's interest as a set of concepts that have a relationship with the domain model concepts. We use these relationships to personalise the search and make recommendations for the user. CH\_INTEREST vocabulary has been created to express the information about a user's interest and their related context, in a structured manner, contained in a single dynamic, extended FOAF profile.

The FOAF ontology alone is not sufficient for our purposes - it needs to be extended with the missing fields and designators, i.e., the missing fields are catered for by the extended vocabulary and are further expressed by mapping to concepts in ontologies like DBpedia, SKOS and GeoName. FOAF profiles already provide an interest property foaf: interest to express a person's interest in a certain subject area. We aim to expand this property further. However, the most important question is about the classification of values for interest identification. In current practice a distributed list of URLs of actual Web sites describing these interests is normally used for this purpose. While in the Facebook case interests are described through a link to corresponding Facebook pages. This approach is not practical if we look at the ultimate goal of the interest profiling done in this research i.e., implementation of a self-sustainable and open to all interest information system. To provide a list of all possible property values by searching the Web and identifying all possible URLs which might be used to describe a user's interest is therefore not an appropriate solution. We propose to use DBpedia classification system as the interest classification in a similar way as WordNet is used for describing nouns.

The CH\_INTEREST extension is designed to support the interest of the user; it is a dynamic profile model, i.e. it keeps on updating itself with new information feeds



from the user's public information whenever the user logs on to the system. The user can view his/her Cheri interest profile and make amendments to it, e.g., if a user feels that the system has misinterpreted his/her interests by pointing an interest term from his/her profile to a wrong concept, the user can remove such concepts from the profile. The system shows why it has deduced that the user might be interested in a certain concept. For example the system found 'Lord of the Rings' in a user's favourite books section from Facebook and deduced that the user might be interested in fiction novels. Then the system will put fiction novels as an interest term in the user's Cheri interest profile and will mention 'Lord of the Rings' next to it as the reason. This portion of the user profile is editable and hence the user can add/remove the interests he/she does not agree with.

### **Ontology Model**

There are two ontologies that are developed for our system; user interest ontology which we call CH\_INTEREST (an extension of the FOAF ontology) and a domain ontology, which we call CH\_ONTO.

*Determining Scope for our Ontologies:* The domain covered by the CH\_INTEREST extension of the FOAF vocabulary is the interest captured by the user's social data and therefore is designed on the basis of the major categorizations of the actual data obtained. The CH\_ONTO ontology is comprised of a lightweight representation of the V&A Data (more specifically V&A objects) shared through the V&A collection API.

### **CH\_INTEREST**

In order to support personalization in the Cheri system, we need to take into account users of this domain. The users of Cheri are general social Web users. These users have different interests when they use the Cheri System. In order to adapt to these needs, we represent the user profiles in the FOAF format. However as discussed above the foaf: interest entity of the FOAF vocabulary is not sufficient to describe and elaborate a user's interest, so an extension is much needed to fulfil our requirements of user interest representation. The extension is provided as an interest vocabulary that we named CH\_INTEREST. The CH\_INTEREST is used to model user interests and their attributes (which are added to describe and detail the user's interest through LOD), as a set of concepts that may have relationships to domain model concepts. We used these relationships for the purpose of personalization.

CH\_INETERST itself is composed of class vocabulary and a set of property vocabularies as discussed below.

### Concepts Used in the CH\_INTEREST

CH\_INTEREST provides a simplistic approach to extending the FOAF ontology. It extends the FOAF: interest property by providing more detailed vocabulary related to the user interest. The vocabulary has only one entity (class) named ch\_interest. This is aimed at expressing the interest of the user gathered from their social profile in an elaborate manner in the context of cultural heritage.

ch\_interest- The class that is used to represent the user's interest.

```
<owl:Class rdf:ID="ch_interest:interest">
<rdfs:subClassOf rdf:resource="&owl;Thing"/>
</owl:Class>
....
```

The ch\_interest concept is used to create relationships between the user and the domain model concepts. Hence, depending on different interest types, diverse relationships exist between the user and the domain entities.

### Set of Properties used in CH\_INTEREST

Table 5.2.3.2 summaries the set of properties used to connect the concepts used in CH\_INTEREST to the domain ontology concepts.

Table 5.2.3.2: Properties of the CH\_INTEREST

Property Name	Description
hasWeight	Defines the weight (i.e. importance) of the interest. This is calculated using the IWCA algorithm described in the next chapter.
hasCatagory	Defines the genre of the user interest
hasSubClass	Defines the sub class relation of the user interest
isRelatedTo	Defines the relation of the user interest to the V&A object
isStronglyRelatedTo	Defines the weighed relation of interest with V&A object
hasRelatedPerson	Defines the people related to the user interest. This may refer to e.g. a well-known player in the case of a sport, writer in the case of a book, or artist in the case of painting.

hasRelatedResources	Defines useful links with the LOD resources
hasRelatedImage	Defines links to the relate images from the LOD. For example it may refer to related images from flickr.

## CH\_ONTO

Using the Cheri system, users can navigate the information space using the CH\_ONTO ontology hierarchy, ontological relationships and dynamically generated hyperlinks. Besides, this information is personalised according to user interest profiles (using background knowledge e.g., user-location, and interests). CH\_ONTO is implemented with reusable components and can be adapted to other ontology domains with a low cost. For supporting navigation, concepts from the ontology are presented as filters along each recommended item (see Figure 6.2.4.2). When a user clicks onto a concept presented to them from the Cheri ontology hierarchy, an ontology-based search query is auto generated and the presented set of recommended art work is modified to present the related artwork corresponding to the original search yet modified with the instances of the selected class (Figure 6.2.4.2). If the user is logged into the Cheri recommender, Cheri also adapts the information to the user. For example, according to the interests of the user, information resources are weighed and ordered using the IRWA Algorithm presented in Section 6.2.6. An ontology based refinement over the ordered recommendation list of Artwork can be done by selecting different properties in the ontology, presented as filters. Whenever a user clicks on an instance from the main recommendation panel, more information is shown in the detailed view visible by clicking the drop down panels named ‘Web recommendations’ (see Figure 6.2.4.1), dynamically generated recommendation links to related instances, using LOD resources such as DBpedia and Flickr which are presented in addition to the links coming from ontology. Furthermore, recommendation links are annotated with visual cues depending on their relevancy to the user’s profile (Figure 6.2.4.4 Geo Results).

### Concepts Used in the CH\_ONTO

CH\_ONTO is an ontology that has two entities: Cheri\_User and vam\_Object. The vam\_Object entity is used to classify different V&A objects according to their attributes. Each user is identified as an instance of the Cheri\_User entity. The

Cheri\_User entity is used to create relationships between the user interest and the domain model concepts. Hence, depending on different user interests, diverse relationships exist between the user and the domain entities. For example, users can add relevant interest topics into their profiles and the system can generate different weights depending on their interests to be assigned to the relevant object.

Table: 5.2.3.3: Set of Properties used in CH\_ONTO

Property Name	Description
hasUser	Defines the user of the Cheri system.
hasCurrentLocation	Defines the current location of the user from which they are access the Cheri application. Will have longitude and latitude values.
hasInterest	Defines the interest of the user. Defines the URI of the user's interest. It can take values of the CHERI_INTEREST instances.
hasObjectType	Defines the type of the vam_Object instance in the vam object hierarchy.
hasTechnique	Defines the technique used to create an instance of vam_object.
hasOriginatedFrom	Defines where a particular vam-Object instance has originated from. Will have a location value defined by a URI.
hasArtist	Defines the creator of an instance of vam_Object.
hasDescription	Defines what is the vam_Object instance.
hasPeriod	Defines the Time period when an instance of the vam_Object was created. Will have a date-time value
hasWeight	Defines the weight (i.e. importance) of the interest in CH_INTEREST.
isRelatedTo	Defines the relation of a V&A object instance with a user interest instance

#### 5.2.4 Query Refinement and Recommending

The user's interest profile informs a recommender system that is an open as well as a closed corpus recommender.

*Closed corpus* as, the Recommender system queries a repository of cultural heritage data of the V&A museum.

*Open corpus* as, one of the objectives of the project is supporting the recommendation mechanisms for the open semantic Web. Thus the system also provides recommendations by querying the LOD over the Web and suggests interesting things related to the user interest that are present on the open linked Web.

### ***Query Refinement:***

There are several methods in recommender systems to obtain user feedback on results for refining the query, filtering and improving the recommendation quality. Some of the most common ones include soft feedback (i.e., like, dislike), ranking and hard feedback (e.g., bread-crumbs, undo). Several filtering mechanisms like adaptive and exploratory path retrieval (AXP) (Cao, 2009), Lazy-DFA (Deterministic Finite Automata) (Chen and Wong, 2004), XFilter (Altinel, 2000) and XTier (Chan, 2002) are also used for the purpose of query refinement. Query refinement through relevance has been well studied in the field of information retrieval (Rocchio, 1971; Ruthven and Lalmas, 2003) and multimedia (Li, 2001).

*Vector Model based techniques:* Most of the existing algorithms rely on the vector model that describes the composition of the data or the query in terms of its constituent features (such as colour, edge, or keyword). The vector is especially suitable for supporting feedback, because the user feedback can be used in both ways to, (a) move the initial query vector in the vector space in a way that better represents the user intentions or (b) to re-assess the significance of the features so that the query better reflects the user's feedback. However, not all data can be easily mapped to a feature space; this is especially true for data with complex structures (such as graph and tree structured data).

*Ranking based techniques:* Ranked query processing techniques to identify the set of results to be shown to the user include Nearest Neighbours (Roussopoulos and Kelley 1995), top-K ranked joins (Qi, Candan and Sapino, 2007; Tao, et al., 2007; Kim and Candan, 2009), and skylines (Börzsönyi, et al., 2001; Khalefa, et al., 2008; Zang and Alhajj, 2010) to name a few.

*Clustering Query Refinements by User Intent*: User interest measures based on the user's document clicks and session co-occurrence information are used as the bases of a user interest measure in these methods. Related queries in the search systems are typically mined from the query logs by finding other queries that co-occur in sessions with the original query. Query refinements are then obtained by finding the queries that are most likely to follow the original query in the user session.

Our Approach for Query Refinement:

Our system does two sorts of query refinements; Ranking based query refinement and Ontology based query refinement. The details of the ranking mechanisms adopted by the Cheri system are discussed in chapter 6.

Ontology based query refinement is achieved by designing a *term refinement service* which suggests and ranks relations, which connect the initial seed concepts with the perspective concept in the domain, and then suggests the query refinement based on the original interest data and the associated ontology concepts.

The *Query support and Recommending Model* (Figure 5.2.1) handles requests from the users. It forwards its requests to the *ontology support engine* to infer the implicit relations between instances based on the knowledge from the user interest profile model and helps expand the query. The related concepts are then forwarded back to the query support which generates the appropriate SPARQL query to query the end resources available to our system including the LOD. The recommended results are then displayed to the user.

### 5.3 Knowledge Resources

This section describes V&A Museum London's online collection database and the freely available linked open data on the Web as our data query sources of choice.

#### **Victoria and Albert Museum Online Collection**

The Victoria and Albert museum online has a working database called *Search the Collections* which comprises of over a million accessible records through the Victoria and Albert museum API. We have chosen this online collection as our CH query

resource. There are three different types of records available; ones that have high quality images, ones that have more basic images and then there are those that have no images. For our research we are considering just those records that have high quality images with detailed information.

As V&A is constantly updating the database. One cannot cache or store any content returned by the V&A API for more than four weeks according to the terms of use of the API. It is also stated in the V&A API's terms of use that "If you display images in your website or application you must use the image URI returned by our API rather than create a copy on your local web server". So the Cheri system retrieves the images directly from the V&A Web collection at run time, no local copies of the images are made only reference URI's are added to the knowledge base whenever necessary.

The API query responses can either be obtained in serialised JavaScript format (json) or as XML. However the json format is found to have more detailed information about the objects. So json is used as our prefer response format.

## **Open Linked Data**

Apart from recommending Cultural Heritage related results, it is useful to provide the user with results from the open Web. In this regard, recommendations from open link data resources like DBpedia and flickr are provided to the user.

In 2010, the Linked Open Data project counted 13 billion triples of LD on the Web (Möller, et al., 2010). The ability to move between data linked in such a way opens up the possibility of exposing data on the Web and being able to access it from any application. The advantage of this is that when data from other sources is accessed, following the links gives the information user access to a contextualisation of the data, or to more information that can be exploited about the subject. If the data retrieved is also linked, then following those links gives access to more information, and so on. Linking data therefore allows the creation of an extremely rich context for an inquiry which furthermore can be interrogated automatically.

Keeping these advantages in mind our project makes full use of the related linked open data resources available on line and also aims to link all data linked and produced by the system according to the LD principles.

## 5.4 Conclusion

This work will prove useful for *exploratory* search refinements, and personalised recommender services. The implementation follows the belief that the set of formal and conceptual technologies developed thus far should be used instead of developing localised ontologies that will have little reuse. Keeping this in mind, well established standards in the field of Semantic and Social Web like FOAF and LD are used.

In this chapter, we have introduced the methodology and development of our user interest profiling and recommending system *Cheri*, which gathers user data from user's SNS profiles dynamically and generates a portable interest-profile for the user. Based on the interest profile thus generated the *Cheri* system recommends art work and related links from the linked open data to the user. Our experiments to evaluate the different research questions and hypotheses as discussed in chapter 1 and 3 are presented in the next two chapters.

Once we have gathered the data from Facebook, either via the application or the website, we aimed to organise them inside *Cheri* using the proposed *Cheri* architecture according to the *user interest semantic modeling technique* discussed in this chapter, to map the various relationships between the user data and the retrieved cultural heritage information. In addition, *Cheri* provides the necessary interfaces for any new provider to map its own data model onto the *Cheri* model or to extend the ontology by defining its own unique attributes. During the evaluation where we test data extraction through Facebook graph and relevant information for the V&A collection to be suggested to the user, we also evaluate the *Cheri* ontologies described in this chapter by exposing the ontological concepts to the user along with each suggested information in their interest profile and evaluate whether our users would understand the value added by the recommendations and their data.

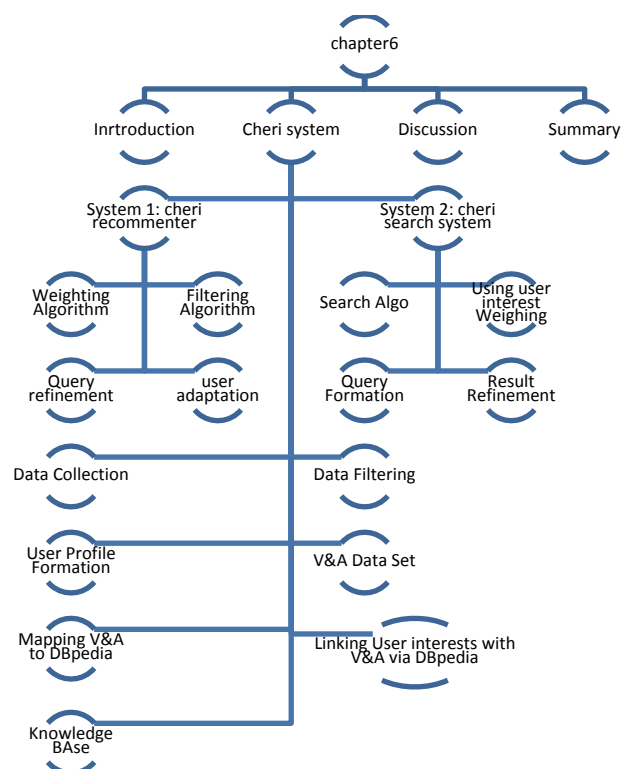
Once the user has decided to export their information into *Cheri*, the system filters and cleans the user data into meaningful concepts and assigns a URI to each concept, via the *Automatic link generation service* i.e., our DBpedia\_URI allocation tool, which is a REST architecture based service and was presented in Section 5.2.2. Once the URIs' have been assigned the service notifies the *Cheri* system that the *Cheri* System can now perform a get request on these URIs to retrieve useful attribute values



to be stored or presented to the user. In the next chapter we will see in Fig 6.2.1 an example. In order for the system to work properly, *Cheri* has to assign URIs automatically to the exported user information and inform the system that by performing GET requests to the user interest URIs, the system can now retrieve the values of the URIs. Generating URIs to display the interest' information along with adopting the REST protocol for allowing the *Cheri* system to perform the GET requests on these URIs worked as expected. The DBpedia\_URI tool did not fail the challenge of generating URIs in real time, and the action of performing a GET request on the generated URI. Another noteworthy attribute of this system is that we did not simulate the V&A environment inside *Cheri*, but instead used the real V&A Web collection resource to validate our proposed communication protocol. The generated URI for the user interest showing links to V&A data and the information and links that the URI was exposing were displayed without any problems.

## Chapter 6

# Cheri System Design



Chapter 6 topic hierarchy

### 6.1 Introduction

This chapter details the design implemented in order to develop the *Cheri* system. The general overview of the two *Cheri* systems (search and recommender system), has been given. The different algorithms that are developed to achieve a working model of these systems have been explained in detail. The systems are developed in such a way that the user can easily evaluate their functions. These evaluations and their results are discussed in chapter 7.

## 6.2 *Cheri* system- Discover, Retrieve and Recommend

The *Cheri* system is a user interest capturing, profile generating and art recommending system designed to make the Cultural Heritage domain more reachable to the general Web user. The interest profile generated through *Cheri* is mapped through LOD standards which make it reusable across the Web as well as machine readable. The interest profile is however layered with a mapping layer to provide the domain specific knowledge provided through the CH\_Ontology. The system uses the interest profile to recommend artwork from the art collection of the Victoria and Albert Museum in London that currently contains over a million records (V&A Search the collection, 2011), as well as open source information from DBpedia and the Web.

Figure 6.1 gives an overview of the system and its various components. The *Cheri* project is designed as a two phase system, the first system being an art recommender and a general Web suggestion system. The second phase of the project is an extension to the first and it harvests its capabilities into a search system. The other details of these systems, the approach in implementing them and the resulting outputs are discussed hereafter.

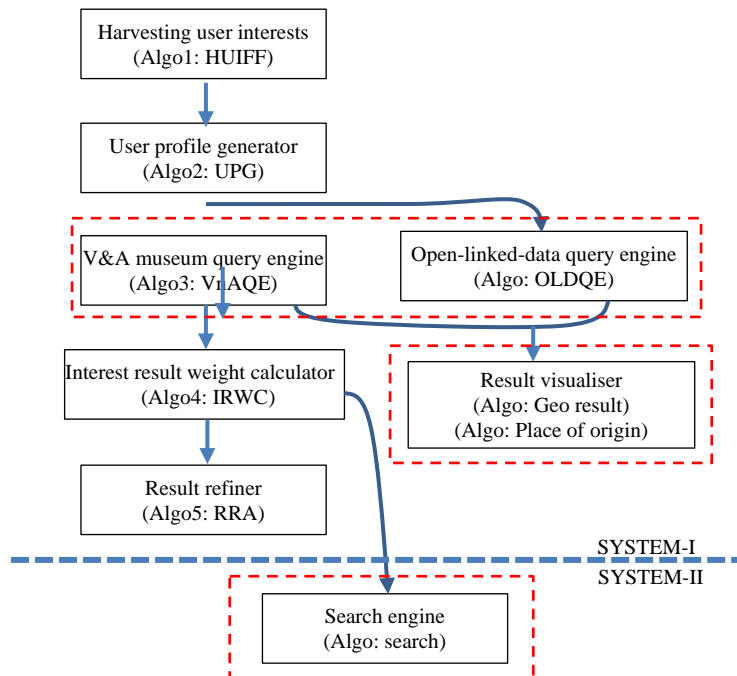


Figure 6.1: Overview and relation of the two systems

### 6.2.1 Data Collection

As discussed earlier in chapter 5, section 5.2.1, Facebook has been used as an example of an SNS for data collection to demonstrate that the SNSs can be used to capture user interest. As the general data collection methodology has already been discussed in chapter 5, the Facebook specific data collection techniques, adapted for the *Cheri* system, are discussed here. Also discussed in this section are the following two ways that are designed to help the user to collect data from Facebook:

- Retrieving User Data via the *Cheri* Facebook Application
- Retrieving User Data via Facebook Graph

#### **Retrieving User Data via the *Cheri* Facebook Application**

A Facebook application, was developed by using the Facebook platform. The implemented Facebook application can be found at (<http://apps.facebook.com/Cheriwelcome/>). The emphasis was not just to evaluate the interoperability while building this application but also to target the user engagement with the *Cheri* environment and the user experience (UX) while using the *Cheri* Application. Later, this base application will be used for a set of user evaluations which will be discussed in chapter 7. Initially, the features for the interoperability were developed to see if the *Cheri* website and the Facebook platform were compatible, which were later enriched with other features that allow users to engage further with the *Cheri* application.

The way in which the user privacy and user data integrity has been addressed is that when the user first adds the *Cheri* application, consent to use the user's interest data is taken informing the user about the information that will be gathered from their Facebook profile and asking whether they allow the application to use this information or not (as shown in Figure 6.2.1(a)). Only with the approval of the user, the *Cheri* application proceeds to extract the desired user interest information. At any time during the evaluation the user can choose to quit the experiment by removing the application from their facebook profile (as shown in Figure 6.2.1(c)). This automatically removes all their data from the *Cheri* data store as well.

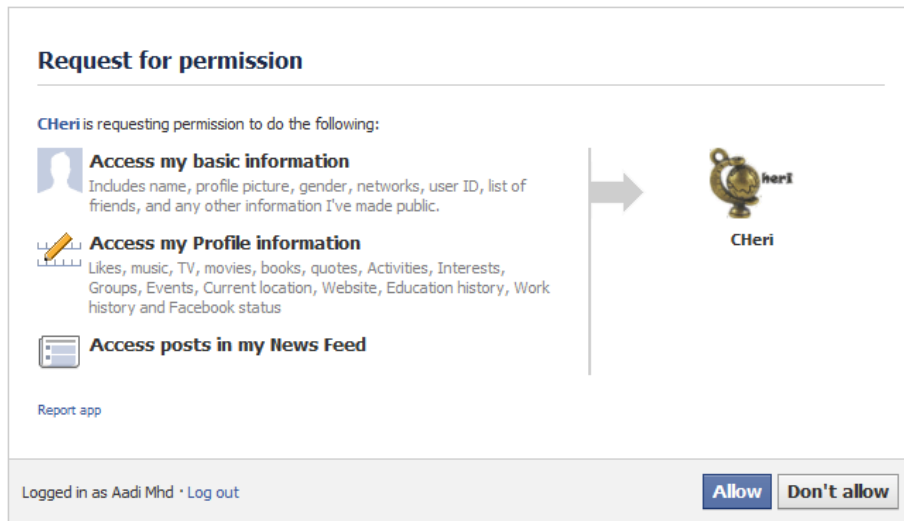


Figure 6.2.1 (a): Allowing *Cheri* to extract user interests from facebook.

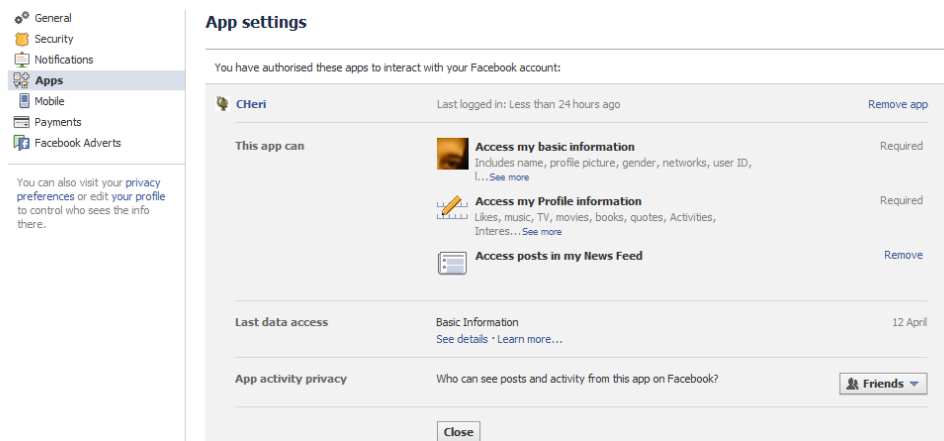


Figure 6.2.1 (b): Options for users to give access of different parts to their profile to *Cheri* application.

## App settings

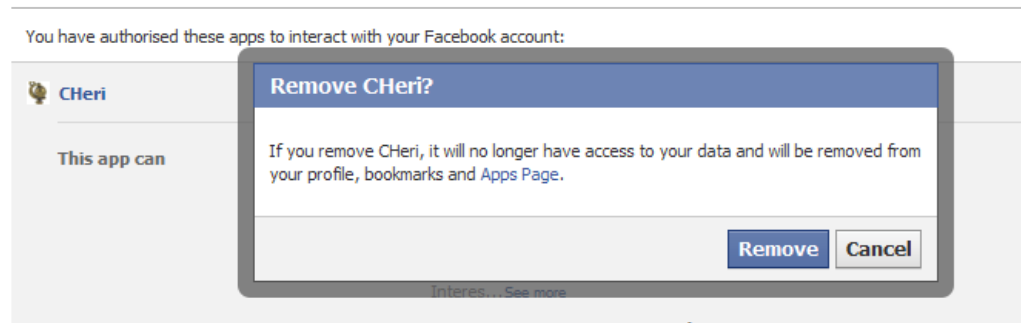


Figure 6.2.1(c): Removing *Cheri* from user profile.

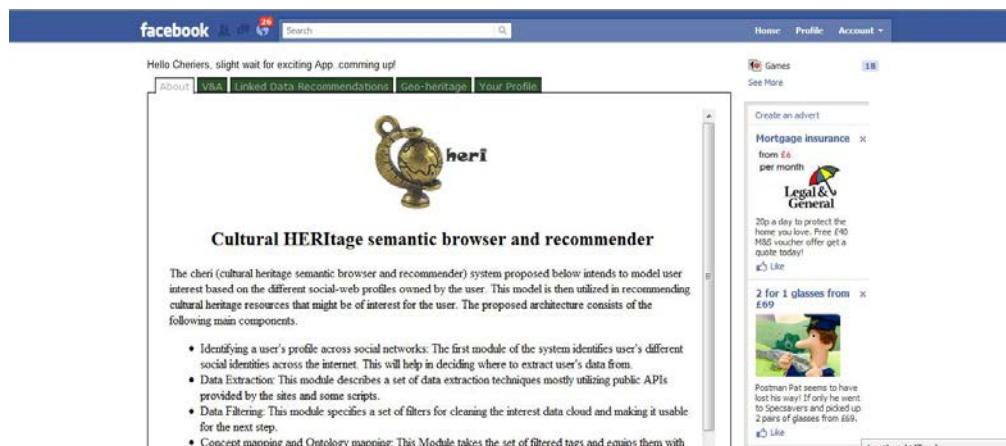


Figure 6.2.2: *Cheri* Welcome Page with *Cheri* Project introduction for the user.

The first thing the user comes across when visiting the application is the welcome page, which explains the purpose of this application (as shown in Figure 6.2.2). The welcome screen introduces the *Cheri* model architecture and provides a graphical representation of the *Cheri* vision. It provides all navigation options as tabs which are explained below.

The *Cheri* facebook application consists of five sections which are presented as tabs and are explained below. Each tab is comprised of different services of the *Cheri* recommender system which are evaluated in chapter 7.

### 1- **About** tab (*Cheri* Model):

The About tab introduces the *Cheri* research project as shown in Figure 6.2.2. This helps the user to better comprehend the proposed architecture. It also helps the user understand the purpose and objectives of the *Cheri* project. When the user first opens the application and authorises the usage of the application, the user data displayed is copied into the *Cheri* knowledge base, unless instructed otherwise by the user/owner of the data (Figure 6.2.1(b)).

### 2- **Recommended Art Work** tab (The V&A Chapter):

The *Recommended Artwork* tab as the name indicates suggests related artwork to the user, based on the interests gathered by the *Cheri* system. Currently, it makes the recommendations from the Victoria & Albert museum online collection only (as shown in Figure 6.2.3). However, this system is platform independent and is capable of integrating other resources as they become available. The evaluation details of this section can be found in chapter 7.

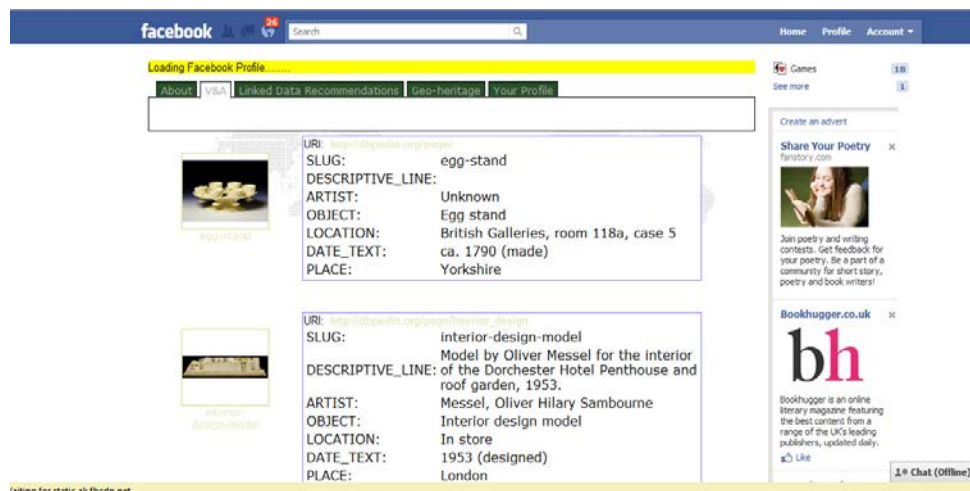


Figure 6.2.3: V&A recommendations

### 3- **Linked Data Recommendations** tab

This option is provided for *Cheri* users to evaluate the recommended resources from the open-linked-data clouds. This mostly includes information related to the user interest from resources linked to DBpedia but also includes interesting recommendations from some other resources such as flickr.

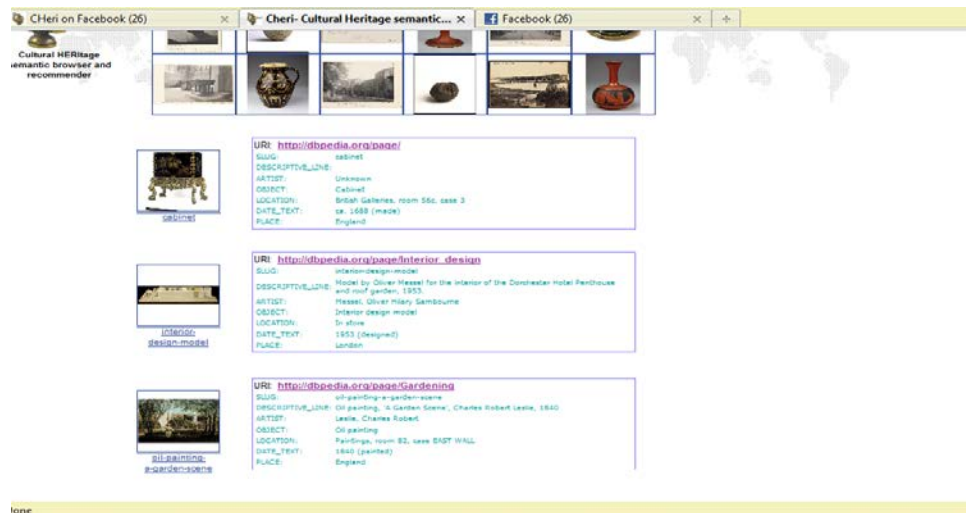


Figure 6.2.4: Open-Linked-Data recommendations

#### 4- Geo-Heritage tab

The Geo-Heritage tab invites users to view their interest related artwork depending on their geographical location. This tests the idea of a walking pervasive museum and the geo results are viewable in two ways to the user. In the first view the user can explore the recommended artwork on a map. The artwork is placed on the map based on its place of origin as can be seen in Figure 6.2.4.3. In the second case the user can choose to view the artwork based on his/her current location. This view presents the user with artwork recommendations based on the location from which the user is currently using the *Cheri* application (as shown in Figure 6.4.5). Here the user is in Southampton and the system detects this information automatically and suggests the artwork from the V&A museum that has originated from Southampton and is related to the user's current interests. Note that the figure gives the recommendations as a circle around the user's current location. This is due to the fact that the coordinates given with each piece of artwork in our knowledge base are city coordinates and are not more precise. Therefore all the artwork recommended based on a particular area will have the same coordinates, and therefore will point on a single point on the map. Therefore for ease of exploration we have mapped the suggested artwork around the coordinates rather than on a single point. This limitation of the system is inherent in the location data available with the artwork. The details of the evaluation of this *Cheri* feature can be found in chapter 7.



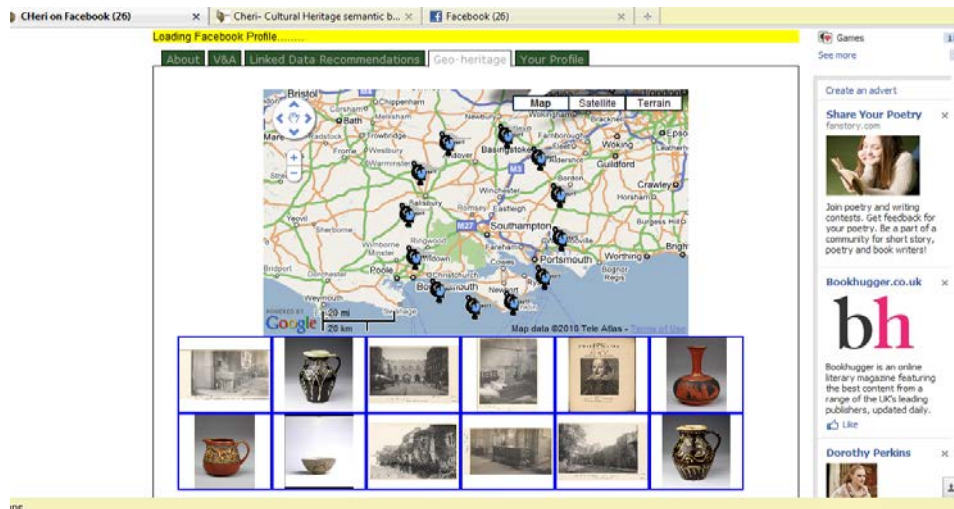


Figure 6.2.5: Geo heritage

## 5- Interest Profile tab

Finally, *Cheri* invites the user to view their interest profile as generated by the *Cheri* system; this is a dynamic profile which updates itself as the user updates their Facebook information. The profile is also portable and reusable as it is generated using the standard FOAF ontology and LD rules. This feature allows a user to carry and reuse their profile in other places on the Web. Details on the evaluation of this *Cheri* feature can be found in chapter 7.

Using the facebook API to implement the *Cheri* facebook application caused some problems from time to time, as Facebook changed the use of some API functions resulting in issues of compatibility of the *Cheri* Application with Facebook, so the code needed modifications to fix these problems.

Worth mentioning are the several social bonuses that were offered by the *Cheri* facebook application. The *Cheri* application helped answer some of the usability questions and helped understanding the end users better (see chapter 7 for usability issues and our observations); the developer and the users alike could take advantage of the facebook invite feature, which could encourages the user to invite their friends to join the *Cheri* facebook application experience. A discussion wall where users can state their own suggestions regarding the *Cheri* experience could also be used. This is a very important resource in testing what works and what doesn't work for the real user and troubleshooting during the developmental stages of the project. In addition to the *Cheri* wall, users can also be given an option to create a discussion topic about any

issue regarding *Cheri*, which we believe can help improve the *Cheri* system in future. Allowing users to discuss *Cheri* and invite friends can only help in promoting and expanding the *Cheri* users' network which would mean more and more people virtually visiting the museums and making use of its resources.

### **Retrieving User's Data via the Cheri Website (using Facebook Graph):**

*A Quick Open Graph History:* Back in 2008, Facebook launched Facebook Connect. Facebook Connect allows people to sign in to an external website using their Facebook account. It was highly successful and within a year, it had 100 million users on Web and Mobile sites. In April 2010, Facebook launched its "Open Graph" API. What this platform does is let developers do more than just connect their site to Facebook. It is a new set of programming tools that helps get information in and out of Facebook. In only one week, the new Open Graph plug-ins were found on 50,000 websites. Initially the *Cheri* website gathered user data using the Facebook connect API but with the introduction of the Open Graph API in April 2010, the *Cheri* data gathering mechanism was upgraded to be compatible with the new Graph API. Open graph meant the opening of facebook data through the modified facebook graph API, with a less restrictive data use policy. That though raised some privacy concerns but proved useful from the developers and research point of view. The system previously used *Facebook Connect* to access publically available user data from Facebook, that had a data storage restriction no more than 24 hours. It gave little time for any complex data processing tasks to be performed over the data. The Open Graph essentially removed this restriction.

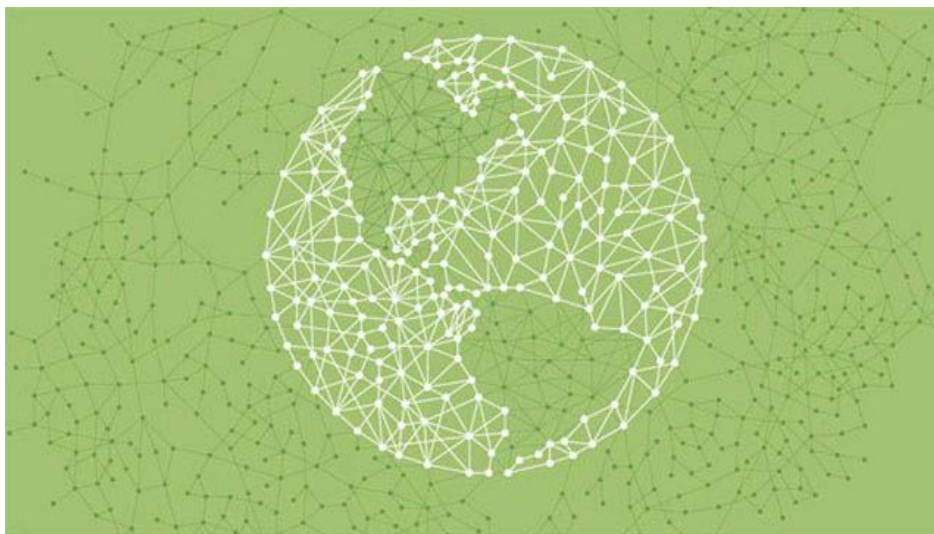


Figure 6.2.7: Facebook Open Graph

Through the *Cheri* website, the second case of user interest gathering was tested. This was a second test for interoperability, the purpose being to observe the issues if any that occurred when a SNS is accessed from outside of its platform. This implementation of *Cheri* can be found at (<http://degas.ecs.soton.ac.uk/Cheri/Cheri-v2.1>).

The *Cheri* website was initially introduced to the users to evaluate three basic aspects of the *Cheri* system, namely: the scrutability, *Cheri*'s artwork recommendation feedback mechanism and the ubiquitous nature of *Cheri* application. However with time and versioning the *Cheri* vision grew, and from a single application that recommended art work to its users it developed into a prototype art recommender and search system so the *Cheri* version 2.0 was tested for an additional 5 aspects, and the 8 different aspects the system was tested for are discussed in detail in the evaluation sections 7.2 and 7.3. During the three evaluations described in chapter 7 of the *Cheri* art recommender and the *Cheri* search system, we also took the opportunity to evaluate the interoperability between the *Cheri* site and the SNS (i.e., facebook) without explicitly stating the intention to our users, to avoid any potential confusion and to keep them focused on the original evaluation task.

The *Cheri* website implementation was initially sectioned into the same 5 components as the Facebook application, but as mentioned earlier with the development of the *Cheri* prototype the sections were further enriched with filtering and exploring capabilities to make it into a complete recommender system. However the basic data collection mechanism remained the same. The details of the *Cheri* prototype system will be further discussed in detail in the following sections. Consider the following screen shorts of the data acquisition process as the user visits the *Cheri* website for the first time.

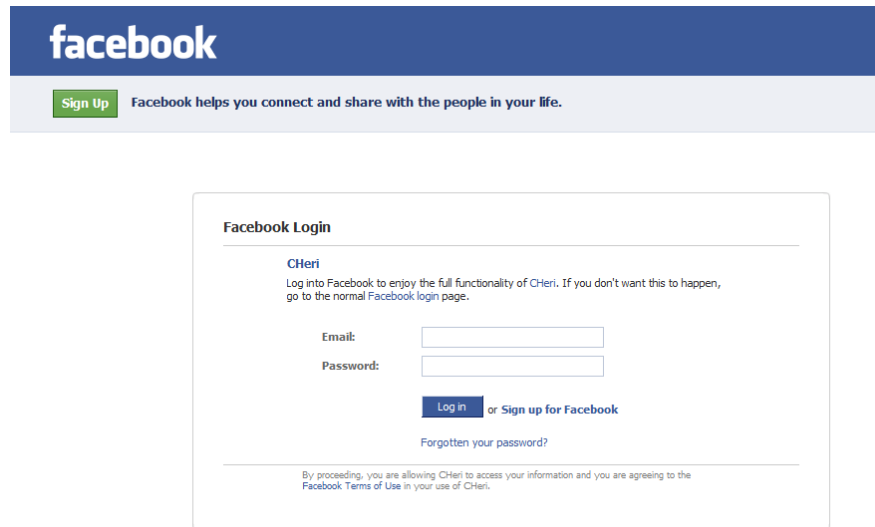


Figure 6.2.8: *Cheri* Facebook login

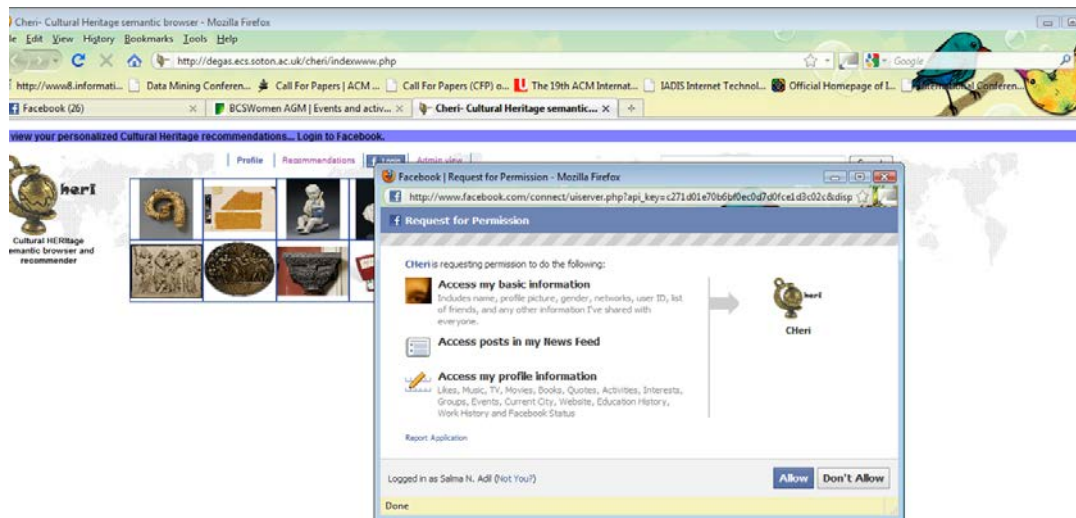


Figure 6.2.9: *Cheri* Website Facebook Data Extraction Request

During the three user evaluations discussed in chapter 7, task 1 asks the users to import the selected contents of their Facebook profile to the *Cheri* system as shown in Figure 6.2.9. The user is required to login to their Facebook account (Figure 6.2.8) and initiate the retrieval of their data from Facebook. After the user's direct consent

has been obtained, *Cheri* retrieves the data and displays them using the *Cheri* model profile. In addition, when the user data were retrieved from Facebook they were copied inside the *Cheri* database, unless the user's instructions were otherwise. In this way we evaluated whether transferring Facebook data from facebook platform to an external website would raise any problems.

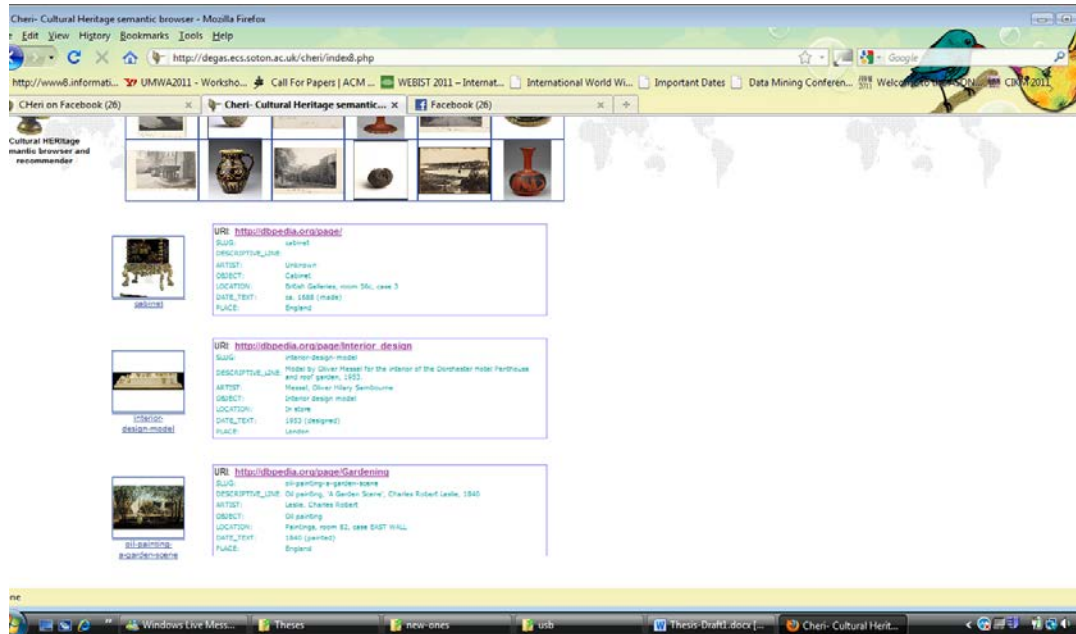


Figure 6.2.12: *Cheri* website recommendation view.

So, using the Facebook graph to transfer user information from the Facebook platform to the *Cheri* website and more specifically to the *Cheri* database was achieved but with some issues that needed to be resolved, as the Facebook platform itself has evolved very quickly during the last 4 years. During the different versions of *Cheri* we kept on modifying the code to accommodate the changes.

### 6.2.2 Data Pre-processing

Though Facebook open graph lifted the time constraints, the majority of user profile data on Facebook was still full of duplicates and ambiguity. As Facebook's Open graph protocol did not resolve or support, object disambiguation or multiple objects on the same page, nor did it apply any mark-up on its pages that could be used directly (till mid-2011, when Facebook provided some mark-up for the user information, which refers to its own object pages) an additional data filtering, semantic processing

and disambiguation service as discussed in the coming sections was needed. Following is an example of some a raw user data obtained through the facebook Graph API in JOSN format.

```
{
  "data": [
    {
      "name": "Flowers",
      "category": "Musician/band",
      "id": "114937881856580",
      "created_time": "2011-07-04T22:41:55+0000"
    },
    {
      "name": "Horses",
      "category": "Animal",
      "id": "111933198826503",
      "created_time": "2011-07-04T22:36:10+0000"
    },
    {
      "name": "Puzzles",
      "category": "Interest",
      "id": "108089669223594",
      "created_time": "2011-07-04T22:25:52+0000"
    },
    {
      "name": "Cars",
      "category": "Interest",
      "id": "110962938928704",
      "created_time": "2011-07-04T22:25:48+0000"
    },
    {
      "name": "How to Train Your Dragon",
      "category": "Movie",
      "id": "96698020019",
      "created_time": "2011-07-04T22:25:48+0000"
    }
  ]
}
```

Figure 6.2.2.1: Raw interest data from facebook

As we can see not much information is available to identify the context or the meaning of most of the interest terms. The need for data filtering and pre-processing is already discussed in greater detail in section 5.2.2 so here we will only discuss the Facebook data and what type of filtering and annotation should be applied on it to make it machine readable and semantically sound.

As such, the use of complex, computationally intensive algorithms on the raw data was prohibited, since this would render the system unusable in terms of responsiveness (if the dataset is incredibly large as is the case in most of the datasets RS systems typically deal with). In addition, the extreme levels of scarcity i.e., the

case if there is not enough information about the user interest in his/her online social profile for the recommender to recommend something relevant, this could have adverse effects on the effectiveness of the framework, which is the reason worth considering while dealing with user interest data from SNS as Facebook.

To illustrate, consider the characteristics of the raw datasets chosen from Facebook.

The following sections provide details of the various data pre-processing steps carried out to overcome such issues. Our algorithm for the data filtering utilises external resources like WordNet to solve the syntactic issues, Wikipedia for synonyms, acronyms and name issues and the general Google search API for resolving misspelling problems. Finally the least frequently occurring terms i.e., the terms which are mentioned less frequently in user profile are removed from the profile and the ones with the highest frequency of occurrence are put forward.

All the data instances that appear in relation to a particular user are first collected and then processed as follows:

### 1. Lexical filtering

A very limited/specific lexical filtering was done on the Facebook data as we needed to keep parts of speech like pronouns and articles that are typically removed during the lexical analysis stage in IR systems. For example, words that are a single character long are removed from the dataset in the general text pre-processing of documents. We needed such terms to make sense of the string as a whole (e.g., *a* in *To Kill a Mocking bird* which is a book a user is interested in). Terms that contain numbers and fall under a global frequency threshold were however discarded.

### 2. Google Spellcheck (Compound nouns and misspellings)

The Google ‘did you mean’ feature provides an excellent way to resolve compound nouns and misspellings. Since this is based on the global frequencies of words on the Web, it is able to resolve common misspellings or abbreviations that would not appear in a standard dictionary. Google is effective for splitting strings consisting of two terms, but is likely to fail for words that consist of more. Since we are also interested in locating the correct concept of the term/word our chosen method of spelling



correction using Google achieves both tasks at one go. The following code snippet (Figure 6.2.2.2) shows how our code concatenates the word/term with the wikiURI as a prelude to searching for the wiki entry of the concept page corresponding to the user interest term.

Figure 6.2.2.2: wiki URI resolving/finding script using Google.

```
function GetdbpediaURI($word){
    $word= str_replace(" ","+",$word);
    $wikiURI="http://www.google.com/search?q=url:wikipedia.org+".$word);
    $content = (file_get_contents($wikiURI));
    $wiki=''.GetBetween22($content,
    '<a href="http://en.wikipedia.org/wiki/', '" class=1>').'';
    return "http://dbpedia.org/page/" . $wiki;
}
function GetBetween22($content,$start,$end){
    $r = explode($start, $content);
    $c=1;
    while($c<=count($r)){
        $wikiCHK=substr($r[$c],0,4);
        if ((isset($r[$c])) and ($wikiCHK!="Wiki")){
            $r = explode($end, $r[$c]);
            return $r[0];
        }
        $c=$c+1;
    }
    return '';
}
```

### 3. DBpedia for resolving concept Disambiguation

The DBpedia disambiguation page is requested for each term if a link to an article matching the term is not found in the first place. If a disambiguation page is obtained, the term is considered to have multiple meanings. In such cases when the user logs into his/her Cheri account to use the application he/she is presented with a list of possible concepts for the ambiguous term the system has identified and an option to choose the right concept for the term.

### 4. WordNet Synonyms

Synonyms are often used to communicate a certain concept. As such WordNet synsets (Miller, 1995), are used to merge together synonymous terms. Moreover, while the filtering of special characters in step 1 does increase tag polysemy; it is required to carry out the Wikipedia look-up step.



Following is the function to find similar words from the WordNet API. The script takes a single term as a string for which synonyms are required.

```
function getSimilarWords($word){  
$word=str_replace(" ", "+", $word);  
$page=curlGET("http://wordnetWeb.princeton.edu/perl/Webwn?s=".$word);
```

The above line retrieves the whole page with synonymous words for the user's interest term stored in the variable, while the function call to GetBetween2 below parses the

```
$simwords=GetBetween2($page, "Webwn?o2=&o0=1&o8=1&o1=1&am  
p;o7=&o5=&o9=&o6=&o3=&o4=&s=", '>').$word;
```

result from the WordNet page and converts it into comma separated list of similar

```
function GetBetween2($content,$start,$end){  
$r = explode($start, $content);  
$c=1;  
$wrddlist="";  
while($c<count($r)){  
$tmp = explode($end, $r[$c]);  
if (strpos( $tmp[0], 'amp;')== false) {  
$wrddlist =$wrddlist. $tmp[0]. ",";  
}  
$c=$c+1;  
}  
return $wrddlist;#,$start,strlen($wrddlist)-1;  
}
```

words. This function GetBetween2 is listed below. The synonyms are not only used to communicate and reinforce the meaning of a given term but are also used in discovering new hidden results during the search and recommendation stage.

### 6.2.3 Linking User interests with V&A via DBpedia and User Profile Formation

A vocabulary expressive enough to be capable of describing every conceivable resource cannot be expected to be readily available. So DBpedia is used as an adequate replacement as discussed in chapter 4.

We generate a user profile which is based on the extended FOAF ontology and populate it with the interest data from the Facebook ID of the user as discussed in chapter 5. The Facebook concepts extracted as user interests are integrated inside FOAF by extending the original FOAF in protégé and accommodating the data into that ontology as an extended FOAF profile (see section 5.2.3 for details). DBpedia acts as a universal vocabulary here to define concepts and as a source to provide a dereferenced URI to the profile concepts and instances. Similarly a V&A related object ontology called CH\_ONTO with some basic required concepts is developed as discussed in chapter 5 and DBpedia here serves the same purpose as before as described in section 5.2.3.

#### 6.2.4 Result visualization

The GUI of the Cheri system allows the user to visualise the results in a number of different ways. The *linked open data recommendations* are visualised as a scrollable list of recommended links and images. The *V&A data visualiser* provides a panel of top six related artwork images for each user interest along with different filters to further explore and discover new related artwork. The *Cheri* recommender also provides two map views of the recommended artwork namely the *Product based* view and the *Active user location based* view. They are described as follows.

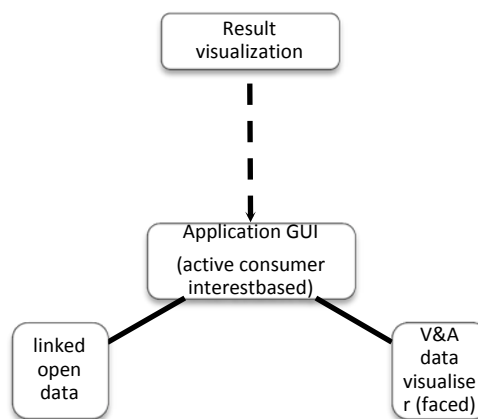


Figure 6.2.4.1: Cheri Active Consumer Interest based Visualizations

*Linked open data recommendation visualiser*: this view is provided under the Web recommendation tab in *Cheri* system. It provides a short description of the user interest that the system has gathered from the user's SNS profile and gives the class/categorization for that particular user interest from DBpedia that it uses as vocabulary. For example here in Figure 6.2.4.2 *Flowers* is the user interest term captured by Cheri from the user's SNS profile and it has been categorised as a member of the class *Garden\_plants*. As the system provides automatic categorization a disambiguation option is provided to the user to correct the concept, using the DBpedia disambiguation list for the page representing that concept.

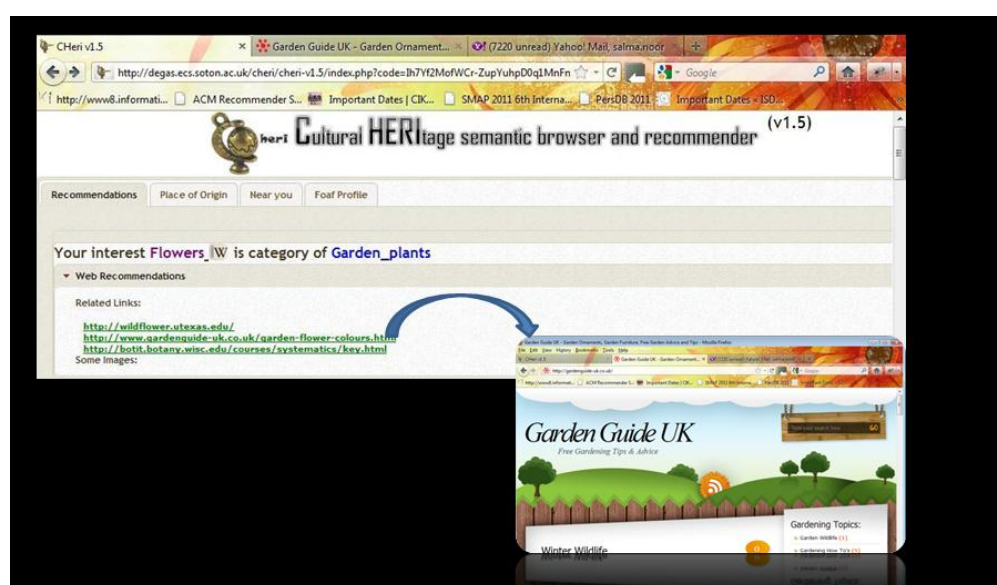


Figure 6.2.4.2: Example of Linked open data recommendation result visualization.

After the concept description a list of related links that might be of interest to the user and a set of images related to the particular interest are presented, which are gathered through the DBpedia related resource links and flickr open image repository API.

*V&A data visualiser*: The V&A Recommendations tab presents the user with a set of related artwork that the system finds to be the most related to a user interest. Moving the cursor over the images reveals some basic information about the image as shown in Figure. 6.2.4.3. We found this was the most suitable way in our case, to present necessary information about a piece of artwork to an inquisitive user without him/her losing the current query flow and without cluttering the screen too much with related information. In addition a set of filters are provided with each artefact to further

modify the search query if the user wants to explore a particular image he/she finds is of interest to them.

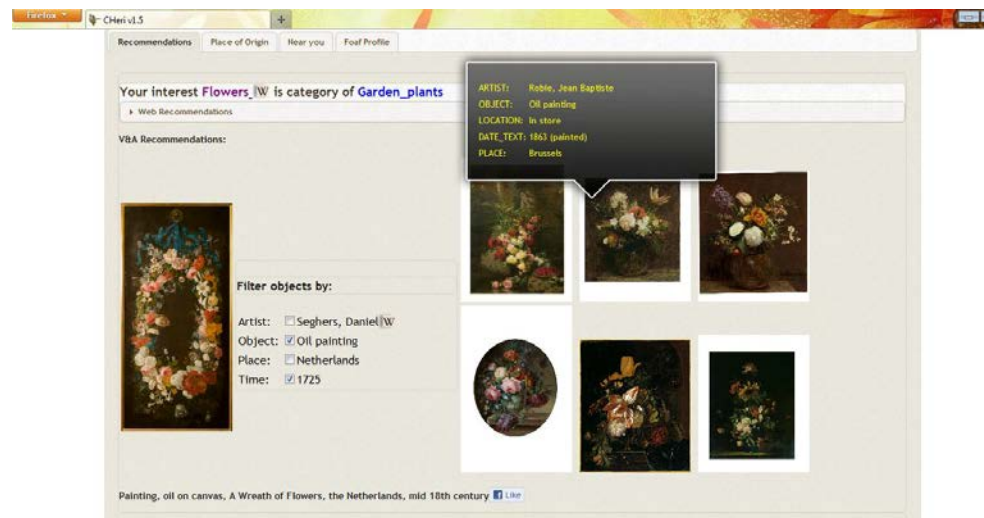


Figure 6.2.4.3: Example of V&A data result visualization.

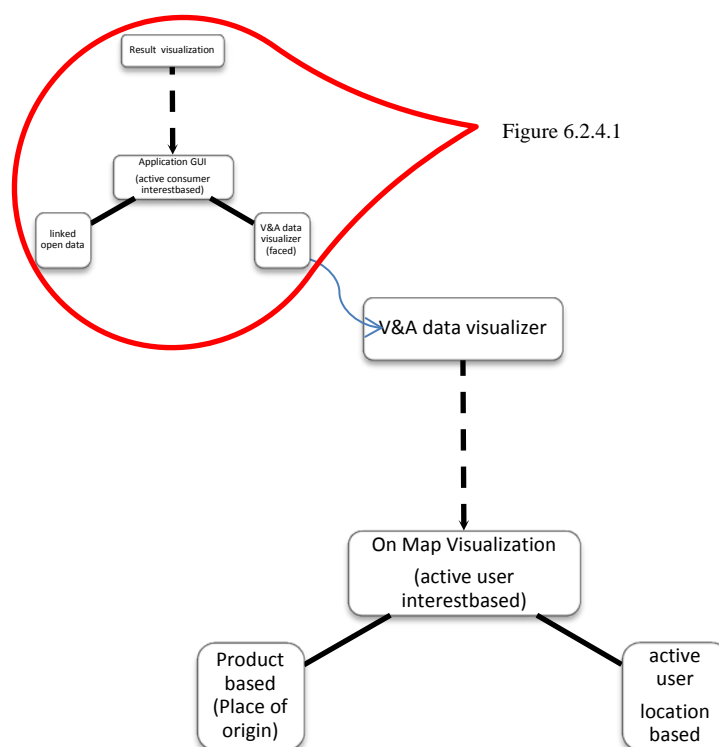


Figure 6.2.4.1

Figure 6.2.4.4: Cheri Active Consumer Map based Result Visualizations (Fig: 6.2.4.1 extended)

The V&A data visualizer present two other ways to view the results/recommendations. Both the methods are map based representations of the results. These options were provided because we intend to introduce the *Cheri* system as a mobile based application in future. And the map based rendering of the artefacts will help us provide the facility of *Cheri* as a walking museum as well as a means of finding the cultural heritage of a new place while visiting it.

*Product Based:* The product based visualization is provided under the *Place of Origin* tab in the *Cheri* system as shown in Figure 6.2.4.5. This option shows each selected artefact at its place of origin, i.e. the place it was made or first discovered. This is an interesting option for a general user and a useful one for a working archaeologist or a historian.



Figure 6.2.4.5: Example of V&A Product based visualization.

*Active user location based:* An *active user* is a user who is currently logged in to use the system. The active user location based recommendations refer to the set of recommendations that are based on the current location of the user in addition to the active user interests. The recommendations are presented under the *near you* tab in the *Cheri* system and represent the artwork from the V&A museum that has originated from or is related to the user's current location, as shown in the example in Figure 6.2.4.6.



Figure 6.2.4.6: Example of V&A active user location based result visualization.

### 6.2.5 Knowledge Base

Distributed information sources, externally-defined models, data portability and powerful APIs are the prime requirements for the next generation of Web applications. Run-time-adaptable applications and the ability to efficiently combine different technologies and formats will more and more be an important factor of success for new Web applications (Nowack, 2011). Scripting languages like PHP have always been a good choice for dynamic environments and integration tasks. PHP helps reduce implementation time and often plays a central role in closing the gap between sophisticated back-end systems and user-friendly front-ends, which makes it well-suited for semantic Web projects too, and therefore was the language of choice in *Cheri* Web application implementation.

This section briefly explores essential PHP based semantic tools, the choices available and our preferred implementation. Several projects to support the semantic Web in PHP and vice versa have emerged ranging from the PHP version of the Repat RDF parser (Argerich, 2002) to battle tested Drupal components and even full blown APIs like ARC and RAP. As we can see in Fig. 6.2.6, only some of them have survived (Bergman, 2011) while there are still parts of the Semantic Web big picture that are not within reach of PHP developers (e.g., Description Logic reasoners). For the rest of

this work we focus only on ARC RDF Classes for PHP (ARC2, 2011) and RAP: RDF API for PHP (RAP, 2008) because those APIs provide support for most of the components of a Semantic Web application (RDF Parser, serializer, RDF Store, Query engine, Inference engine).

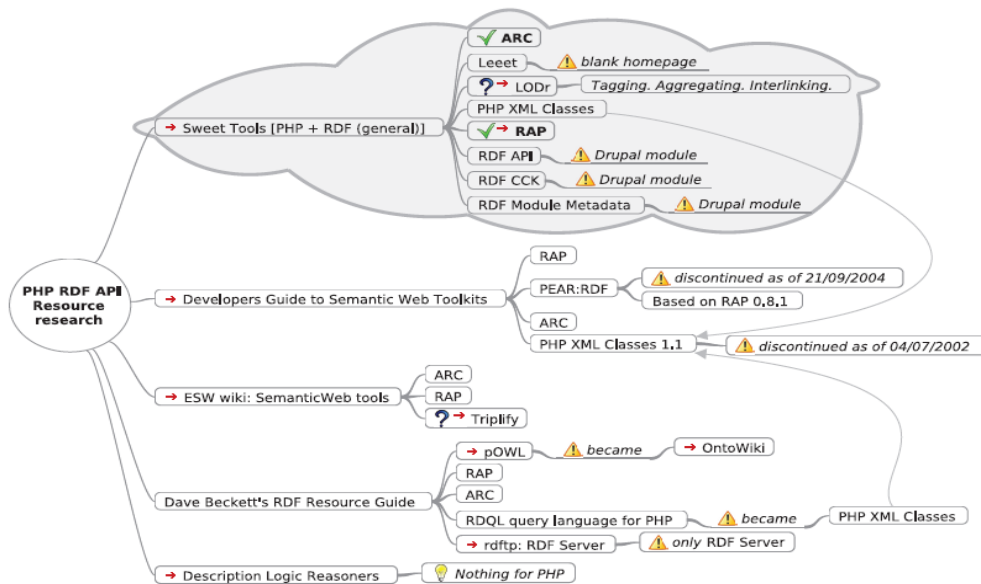


Figure 6.2.5: Searching for modern pure-PHP RDF APIs (Butuc, 2009)

RDF API for PHP, in short RAP, (Chris Bizer: <http://www4.wiwiwss.fu-berlin.de/bizer/rdfapi/>) is a software package for parsing, querying, manipulating, serializing and serving RDF models. It also has an integrated RDF store (quads), SPARQL query engine, SPARQL endpoint support, RDFS and partly OWL inference and a graphical GUI for management of RDF store. ARC (ARC2, 2011) developed by Benjamin Nowack is a flexible RDF system for the semantic Web and PHP practitioners. It is free, open-source, easy to use and runs in most Web server environments. It is already used in several projects. It features several i/o parsers/serializers, an integrated RDF store, SPARQL query engine SPARQL endpoint, some simple inferencing and there are also some plugins. Table 6.2.5 shows a comparison of the two.

The motivation behind ARC was the need for a set of tools that are easy to combine with existing software, in contrast to contemporary toolkits that have non-standard



extensions. ARC started in 2004 as a lightweight RDF system for parsing and serializing RDF/XML files. It later evolved into a more complete framework with storage and query functionality. It realised the need for supporting the already existing Web 2.0 data formats (e.g., microformats, JSON, Atom, RSS2) with a toolkit that was light-weight and optimised for PHP. To achieve this, ARC used object-oriented code for its components and methods, but the processed data structures consisted of simple associative arrays, that led to faster operations and less memory consumption. As of June 2008, ARC's structures are aligned with the Talis platform.

Table 6.2.5: Comparison of Semantic capabilities in RAP and ARC

	Features	ARC	RAP
Solution for RDF storage/ Triple store and Query Engine	In memory and database storage	yes	yes
	Database support	MySQL as RDBMS	Any ADOdb compliant databases
	SPARQL support	SPARQL, SPARQL+ (subset of the SPARQL update)	
	Protocol compliant end-point class can be used for HTTP-based data access as well as a client for remote SPARQL endpoints	yes	yes
RDF Parser/Serializer		has both generic and specific parsers for RDF/XML, Turtle, N-Triples, RSS 2.0, SPOG or HTML and can only serialize in RDF/XML, RDF/JSON, Turtle and N-Triples.	Has parser for RDF/XML, N- Triples, N3, TriX, GRDDL and RSS/ATOM
API-Paradigm		offers Statement-centric and Resource-centric APIs	offer Statement- centric, resource- centric and Ontology-centric APIs
Support for SPARQL queries		yes	yes
Performance		The ARC toolkit proves to be more focused on core tasks (e.g., parsing RDF/XML) thus providing better performance.	
Functionality		Performance comes to the cost of lacking functionality as compared to RAP.	RAP delivers essential reasoning support (e.g., RDFS and some OWL Rules) till a GUI for managing database- backed RDF models, or a graph visualization module, and an in-depth



		documentation for all features
Implementation maturity	ARC enjoys a more effervescent development and community, the latest ARC2 release dated November 2009. As per ARC website 'By 2011, ARC2 had become one of the most-installed RDF libraries. Nevertheless, active code development had to be discontinued due to lack of funds and the inability to efficiently implement the ever-growing stack of RDF specifications.'	RAP has stagnates at version 0.9.6 since February 2006,

After testing, the ARC toolkit proved to be more focused on the core tasks (e.g., parsing RDF/XML) thus providing better performance. Keeping our project requirements in mind we switched to ARC2 (which is an extension of the original ARC platform) from RAP. We chose ARC2 over RAP because it uses a different approach by rewriting SPARQL queries into SQL, so the expensive part of query processing takes place in the highly-optimised database. As time was a big concern we chose ARC. ARC does less validation and operates using lightweight PHP arrays as its data structures instead of objects.

The issue here was that although ARC was fast enough, bugs in its SPARQL support started occurring with larger datasets, MySQL froze when too many UNION patterns were combined; variable bindings contained malformed values in UNION patterns with a different number of variables; language filtering was not fully supported, etc. These drawbacks were overcome by developing our own language filtering and visualization support. We had to implement a more object-oriented wrapper. This resolved our software needs for the current implementation but in general a major shortcoming of the ARC platform is that it works on the RDF level only with some basic inference support and lacks more advanced features of inference support.

#### 6.2.6 System 1: Cheri Recommender

Topics such as data collection, authorization, user interest profile formation, the V&A data set, DBpedia mapping and the ARC knowledge base have already been discussed in chapters 4 and 5 and sections 6.2.1 to 6.2.7. Although these are the major

components of the Cheri recommender system, they are also building blocks for the Cheri search system, so they were discussed beforehand. In this section and the next we will only discuss those features that differentiate the *Cheri* System 1 (the recommender system) from the *Cheri* system 2 (the search system).

In this section we will discuss the rating matrix and our Recommender Algorithm. This section introduces our weighting algorithm IRWA, result visualization, result refinement through user feedback and user adaptation through relevance feedback.

The recommender algorithm performs a series of calculations over a ratings matrix, the rating database or preference database containing the preferences of all users in the system; this is used to generate a top-n list of recommended items.

There are two ways for the user's opinion of an item to be recorded in a recommender: explicitly or implicitly. If the information is recorded implicitly, it is usually gathered at the point of consumption, (e.g. when the item is purchased, but before the user uses the item). Since the opinion of the user is never recorded, this unary information may not be as reliable as explicitly gathered preference information—only the user's initial perceptions are recorded, not their final opinion. If explicit ratings are gathered, the active user returns to the recommender to 'rate' the item, encoding his/her opinion of the item into the ratings matrix.

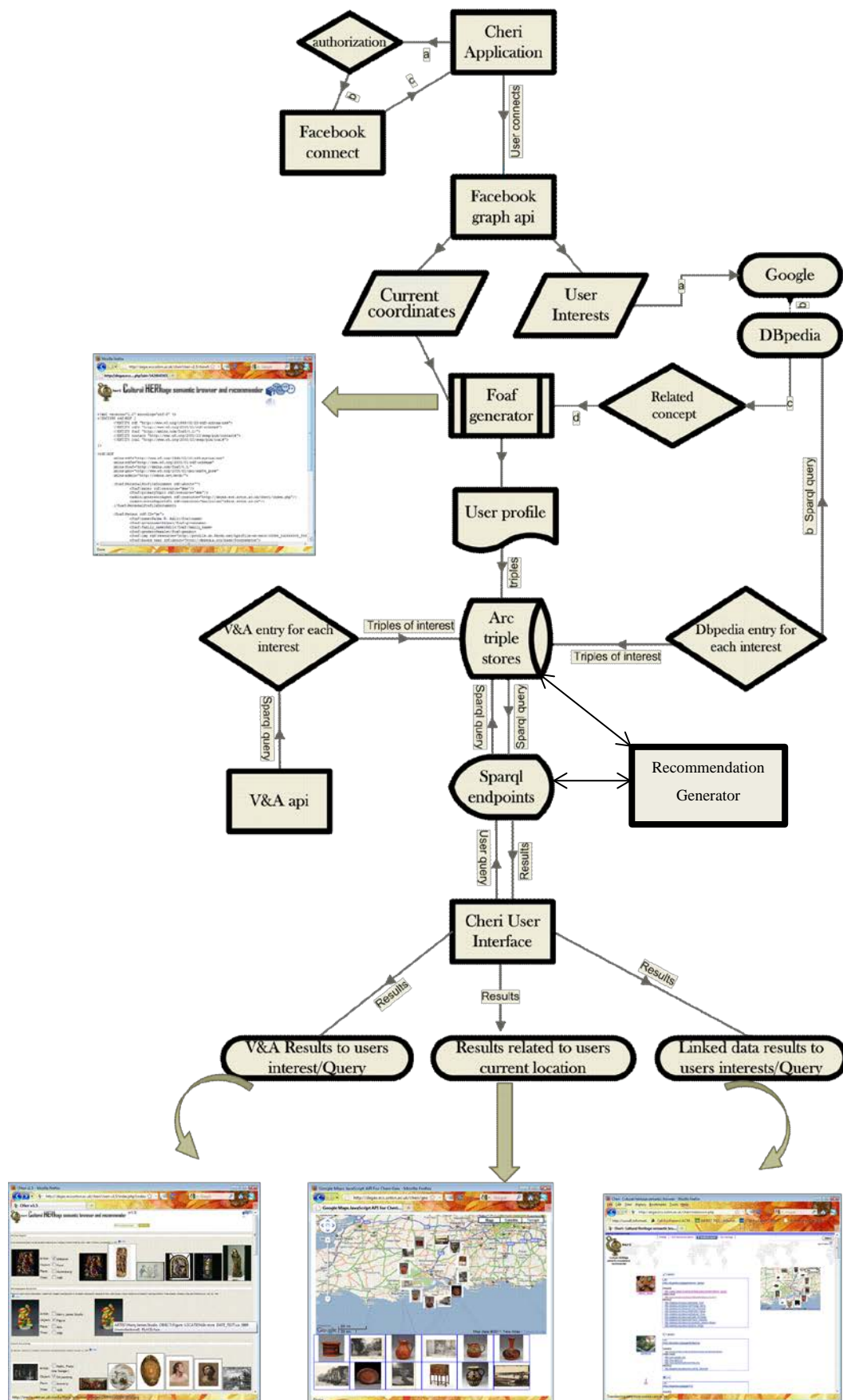


Figure 6.2.6: An overview of the *Cheri* Recommender System

## **The Ratings Matrix**

The ratings matrix is a matrix where the columns of the matrix represent items in the domain, and the rows represent users. Each entry in the matrix represents the opinion that one user has about one item. There are several ways to encode this opinion. In the MovieLens movie recommender, for example, movie opinions are encoded on a 0.5- to 5.0-star scale. Online retailers, on the other hand, might only record whether a user has purchased an item or not as unary ratings. Many cells in the matrix will be empty. That is, not every user will have an opinion on every item and not every item will have been consumed by every user. In practice, a ratings matrix may be stored in a sparse representation or an otherwise compressed format to save space.

The Rating Matrix for the Cheri system follows the general rules of matrix construction but is a modification of the more typical rating matrix in the sense that it is an Interest-rating matrix rather than being just an item rating matrix.

The interest-rating matrix is such a matrix where the rows of the matrix represent the explicitly gathered interests of the user and the columns of the matrix represent items in the domain; in this case artefacts from the V&A museum online art collection. Each entry of the matrix represents the relevance one user-interest has to one item in the collection. This relevance is encoded on a 0.5-to 5.0 scale. Many cells in the matrix would be empty as only a few objects are related to a particular user interest. To save space the matrix is stored in a sparse representation.

The algorithm designed to create the interest-rating matrix is given below and is called IRWC (Interest Result Weight Calculator)

Collaborative filtering and content-based recommenders make very different use of the ratings matrix. In a content-based recommender, the ratings matrix acts as a historical record of the items a user has previously accessed. This information may or not be used to generate future recommendations. That is, content-based recommenders make use of external information gathered from the items themselves to generate recommendations.

---

**Algorithm 1** Interest Result Weight Calculator (IRWC)

---

**Input:** A list of  $U_i$ 's Interests terms (ia)

**Output:** Interests weight matrix for  $U_i$

**ADAPTED-WEIGHTING ALGORITHM** (*array( $U_i$  interest terms)*)

```
1: Connect to V&A JSON API
2: for Each interest term of Array(ia) do
3:     Fetch the first page of V&A queried Object Record
4:     if There is only one page for related Object Records then
5:         Apply Association Analysis
6:         Harvest metadata for V&A objects
7:     else if There is more than one page then
8:         Apply Association Analysis
9:         Harvest metadata of V&A objects page by page
10:    else
11:        break
12:    end if
13: end for
```

---

### **Recommender Algorithm IRWA**

A recommender algorithm has two possible tasks, prediction and recommendation. In the prediction task, the recommender is to determine what value should appear in any given empty position on the matrix. In the recommendation task, the recommender is to determine a list of the empty spaces that the active user will fill in with the highest possible opinion; that is, which items the user will like the best.

When a user asks a recommender about a particular item (e.g. will I enjoy *Starry Nights?*), the recommender makes a prediction about that item. When the recommender generates a top-n list of paintings the active user should see, the recommender makes a recommendation. While similar, the algorithms for generating predictions and recommendations are different. Moreover, the metrics used to evaluate success and the underlying user tasks are different. In this dissertation, we deal exclusively with recommendations.

The Cheri recommender algorithm works in the following steps;

**Pick:** user interests (output of Algo1: HUIFF)

**Extract:** V&A results (output of VnAQ)

**Rank:** V&A results (through IWCA)

The Algorithm IRWA which is our recommender algorithm, calls upon the IWCA to calculate the rating matrix, the cumulative interest term frequency of the V&A object types is calculated on top of the ranking matrix generated by IRWC

**Call Algorithm:** (VnAQ and IRWC )

**input:** (V&A objects)

- 1: Text mine metadata with each V&A object for the interest term and its synonyms
- 2: **for** each interest term match and Calculate the term frequency of all its synonyms in order to find the weight/ ranking of each V&A object according to user interest.
- 3: Assign weight to the object type.
- 4: Find **cumulative ranking** (for all interests)
- 5: determine most frequently occurring V&A objects related to a user's interest based on previous ranking

**Output:** (ranked V&A object types matrix based on user interests)

### **The Recommendation Process and User Adaptation**

Figure 6.2.6 depicts our generic recommendation model. The heart of the model is the Recommendation Generator; all given recommendation calculations contain the recommender and post-processing algorithms.

The flow in the model moves from top to bottom and left to right, starting with the user Interest model and ending with the recommendation list. The Recommendation Generator takes in a representation of the user's current state of knowledge (the user interest profile), the user's Information Need and any specified Settings and Contextual Parameters (e.g., filters the user has selected for an on-going query). The settings and parameters may come from several different sources, including from the user via the user interface, gathered or calculated from user's interest model, from the application hosting the recommender (e.g., Facebook or the browser), or from some internal state of the recommender itself. The Recommender Generator applies our algorithms for a given user interest model and information need to recommend and filter a set of results from a set of data repository at a time (V&A object collection, DBpedia or our local ARC knowledge Base).

### 6.2.7 System 2: Cheri Search System

*Cheri*, the search system works on two algorithms named, IRWA which is a Ranking algorithm and AQOA which is a Search algorithm. These algorithms are explained stepwise in this section.

#### a. Ranking Algorithm IRWA

This algorithm runs when the user logs in to the *Cheri* search system. The stepwise procedure of this algorithm is as follows;

**Step 1.** Retrieving the User Interest: After the user login, accept and permit to access their Facebook profile, the user interests (facebook-interests) are obtained using *HUIFF Algorithm* e.g., a user has mentioned 7 interests, including: *Travel, Food, Swimming, Driving, Painting, Flowers and Horses*.

**Step 2.** Setting the Context for each Interest: For each interest, the spelling check, spelling correction and its context (DBpedia URI e.g., <http://www.dbpedia.org/page/Travel>) is obtained using Wikipedia API. If Wikipedia API fails to retrieve any results, then Google search using following search string “url:wikipedia.org+facebook-interest” is performed and the first Wikipedia URL obtained in the result is parsed as DBpedia URI context for each particular ‘facebook-interest’.

**Step 3.** Storing the User Interest: To avoid the repetition of step 2, next time when the user logs in to the Cheri Search system, the user interest DBpedia URIs obtained in *step 2* are stored as triples, to be retrieved semantically by using a SPARQL query:

```
<facebook-userID><hasInterest><facebook-interest>.  
<facebook-interest><sameAs><user-interest-DBpedia-URI>.
```

... ..

Therefore, the next time the user logs in to the *Cheri* Search system, if the user’s interest already exists in the triple-store, then step 2 will only be performed for the newly added interests.

**Step 4.** Obtaining synonyms of interest: The synonyms of the “*context-term*” in each DBpedia URI (e.g., *Travel* in <http://www.dbpedia.org/page/Travel>) are obtained as JSON object using *WordNet* which is a lexical database for English. e.g., for ‘Travel’ the *WordNet* result is “*travel, traveling, travelling, change of location, locomotion*”.

**Step 5.** Searching, Adding and Auto-predicating objects from the Victoria and Albert Museum (VAM): Using *VAM Rest API*, the context-term (e.g., *Travel*) is searched in VAM database and records of the top 45 VAM-objects (e.g., searching “*Travel*” will give several of “*Photograph, Oil Paintings, Painting, Print, Poster, Drawing, Drawings, ... ..*” objects) will be text-matched, weighted and sorted (in descending order) in such a way that top objects will have the maximum number of occurrences of the context-term and its synonyms. Also, if any VAM-object type repeats in the top 45 results then its weight is cumulated with the initial occurrence. As, for this study, it was decided to recommend only the top five objects to the user, these five objects are predicated as “*<isStronglyRelatedTo>*” and the rest are predicated as “*<isRelatedTo>*” the DBpedia-URI.

For example, in Table 6.2.7.1 below, by searching *Travel*, the 45 results that are obtained, have a total of 28 VAM object types with repetition. The objects are sorted by weight of occurrence (*hasWeight*) for the concept representing the user interest as a DBpedia URI. e.g., the triple for “*Photograph*” that shows a higher weight as compared to “*Painting*” for the interest *Travel* will be stored as

*Photograph <isStronglyRelatedTo><http://www.dbpedia.org/page/Travel>*.

... ..

*Painting <isRelatedTo><http://www.dbpedia.org/page/Travel>*.

... ..

Table 6.2.7.1: Example of sorted weights for “Travel” DBpedia URI, assigned to VAM objects

V&AObject	Auto_Predicate	SearchString	dbpediaInterest	hasWeight
Photograph	<isStronglyRelatedTo>	Travel	<http://dbpedia.org/page/Travel>	78
Board game	<isRelatedTo>	Travel	<http://dbpedia.org/page/Travel>	71
Painting	<isRelatedTo>	Travel	<http://dbpedia.org/page/Travel>	69
Banner	<isRelatedTo>	Travel	<http://dbpedia.org/page/Travel>	26
Roundel	<isRelatedTo>	Travel	<http://dbpedia.org/page/Travel>	26



The reason for choosing 45 objects and not all the search results (which can be more than 1000 objects) is the time constraint, as the whole process of weight assignment, mentioned above, takes around 45 seconds per interest. Some results of weight assignment with Auto-predicating are shown in Table 6.2.7.2 below.

Table 6.2.7.2: Time required for the user in sifting the results, assigning weights to all V&A objects related to a particular user interest (represented through DBpedia URIs) and Auto Predicating.

DBpedia link	Time the Script took for Adding V&A objects and Auto Predicating (sec)
-> <a href="http://dbpedia.org/page/Flowers">http://dbpedia.org/page/Flowers</a>	51.98
-> <a href="http://dbpedia.org/page/Horses">http://dbpedia.org/page/Horses</a>	50.28
-> <a href="http://dbpedia.org/page/Puzzles">http://dbpedia.org/page/Puzzles</a>	37.46
-> <a href="http://dbpedia.org/page/Cars">http://dbpedia.org/page/Cars</a>	34.04
-> <a href="http://dbpedia.org/page/Beach">http://dbpedia.org/page/Beach</a>	44.10
-> <a href="http://dbpedia.org/page/Flying">http://dbpedia.org/page/Flying</a>	43.35
-> <a href="http://dbpedia.org/page/Food">http://dbpedia.org/page/Food</a>	Food has already been added so the script took 0.22
-> <a href="http://dbpedia.org/page/Books">http://dbpedia.org/page/Books</a>	Books has already been added so the script took 0.21
-> <a href="http://dbpedia.org/page/Dining_in">http://dbpedia.org/page/Dining_in</a>	54.94

**Step 6.** Cumulative Frequency of Weighted Objects types for Each Interest: After storing five objects/interests in the triple store, the frequency of each object type (number of times occurrence of an object type for all the interests, altogether) is calculated, sorted in descending order and stored as the Cumulative frequency for each object type.

For example, if the object type “*Photograph*” is found in the results of the top five VAM objects for user interest of “*Travel*”, “*Food*” and “*Painting*”, then the cumulative frequency of “*Photograph*” is 3 and so on.

In this way another list of VAM objects types is obtained and again the top five objects types (e.g., *Photograph*, *Print*, *Poster*, *Drawing*, *Cartoon* and *Oil Painting* with cumulative frequencies of 3,3,3,2,2 and 1, respectively) are selected for the same reason as explained in step 5.

## b. Search Algorithm (AQOA)

This algorithm runs when the user enters a search string to perform a search in the VAM collection. In this process the user search term (e.g., “car”) is optimised using the cumulative frequencies obtained in step 6 of the ranking algorithm, described above. Now when the user enters any search-term, one search result for the search-term is obtained from the VAM for each of the five VAM objects obtained in step 6 and displayed to the user as the *Cheri* Search recommendation.

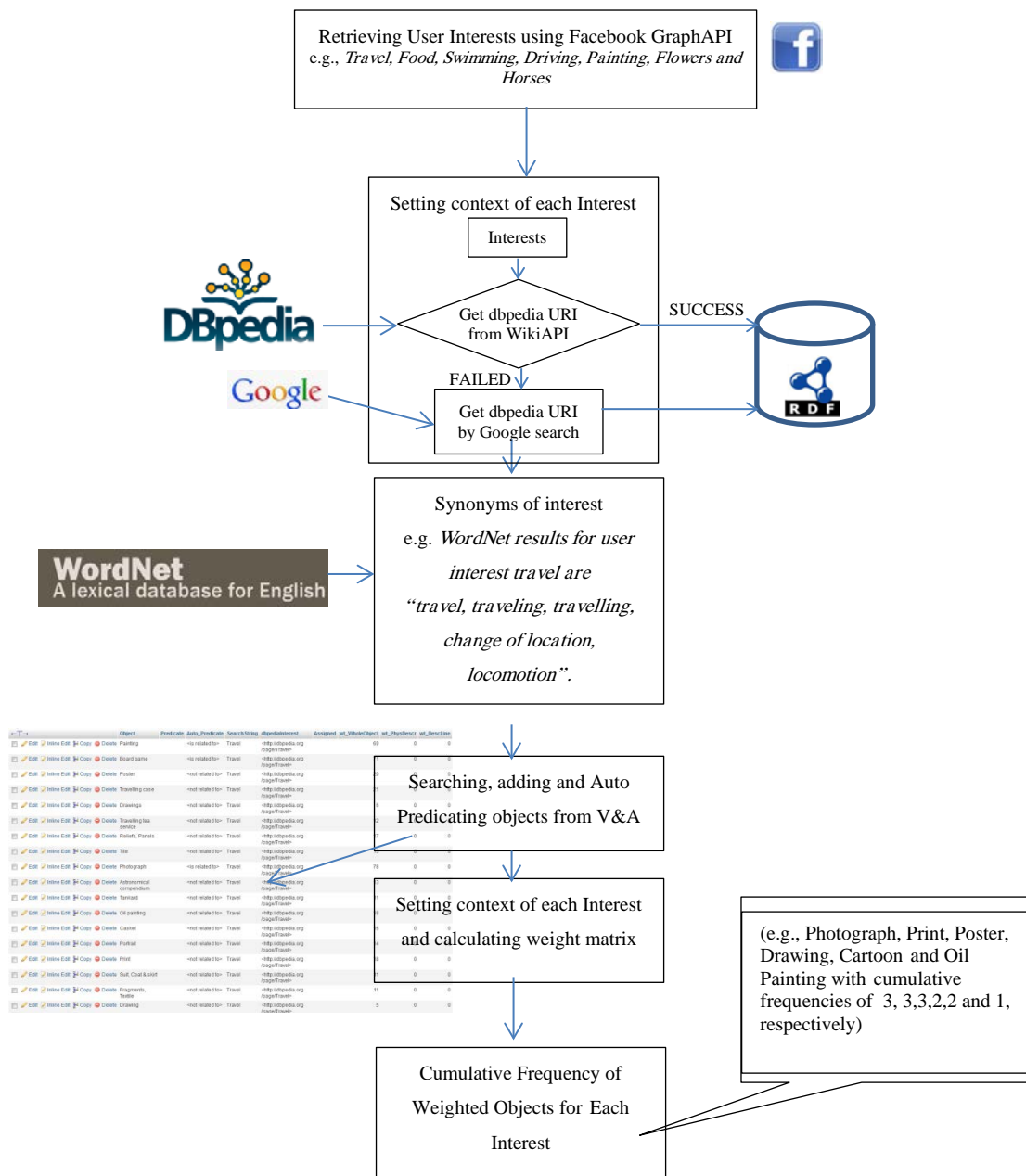


Figure 6.2.7.1: Workflow of Cheri Search System

As mentioned earlier the IRWA Algorithm takes the V&A hits/results for each user interest and assigns a weight for each result. The weight is calculated based on term frequency/synonym frequency (text mining technique). And every new occurring object type is ranked according to this term weighting technique.

The ranking matrix now has the weights for all the *objects* related to the user interest terms. Now another rating matrix that contains the weights for all the *object types* across all the results found for the user interests is calculated. From this ranking the top five most occurring object types for a user across all his/her interests is identified.

Now consider that a user enters a search term. The Cheri system does a query to the V&A museum API against this search term and generates a list of results most related to the user interests based on the object rating matrix. The results are further narrowed down based on the top five object types calculated above. And the results are recommended to the user. Hence using user interest based weighting matrix for V&A query formulation as well as result refinement. As shown in example screen shot below

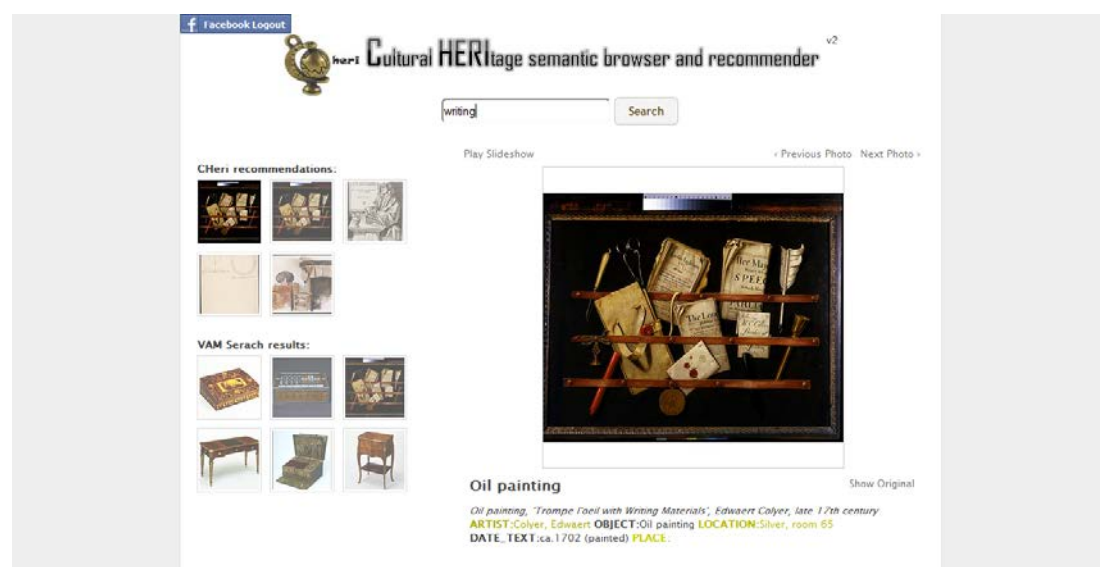


Figure 6.2.7.2: Example Screen-shot of Cheri Search System

This mechanism helps the system in taking the user interest as well as the current search context into consideration while recommending artwork.

## 6.3 Conclusion

The previous chapter set a requirement for modelling user interest semantics to help facility personalised cultural heritage related recommendations to a general social Web user. This chapter proposed a *Cheri* recommender and a *Cheri* search system to achieve the task and describes in detail the components of both the systems, from their conceptualization to implementation, the basic foundational modules and algorithms the two systems rest upon and the functionalities that set them apart. The *Cheri* recommender system does not provide the facility of entering a search term. It rather takes as its input a list of changing user interest terms from their everyday activities on the online social network, generates a user interest profile, and recommends Art work based on that interest profile to the user. The recommended list changes with the changes in user interest on their social network. While the *Cheri* Search system takes in a user query (search term entered by the user) and presents related Art work from the V&A museum online and LOD online to the user based on the search term. The *Cheri* search system however uses the same user profile generation and interest weight calculation method as the *Cheri* recommender to refine the result list from the user query.

Section 6.2 presented the different building blocks of the *Cheri* system whose motto is to discover, retrieve and recommend. The *Cheri* system is an effort towards discovering new ways of bringing the Art and the general web user together; finding and retrieving user interest information for a dynamic and portable user interest profile that could not just be used by the *Cheri* System but could be used independently as a source of user interest information; and recommend user interest related information. Section 6.2 highlighted the difficulties faced in achieving each functionality of the system from section 6.2.1 to section 6.2.7 and the solutions we proposed to tackle the problems faced. The main problems encountered during the process included the shifting of data collection method from facebook connect to facebook graph when the facebook changed its data extraction policy. This required re-coding certain parts of the *Cheri* data extraction mechanism but was achieved successfully. The next problem encountered was the pre-processing required to make the user data usable. This was achieved through a number of steps that involved lexical analysis, Google spelling check mechanism, DBpedia for resolving concept ambiguities and Word Net.

Deciding the right view to visualise data was the next main decision. Here the decision was taken to design Cheri to view data in a number of different ways namely; active consumer interest based view, on map visualization and active user location based view. Designing the appropriate knowledge base came next, keeping our project requirements in mind ARC which is a PHP based semantic tool was chosen and with the release of the next version we shifted the Cheri System to the ARC2.

Section 6.2.1 presents the data collection mechanism used by the Cheri system. The data collection module for the Cheri system is discussed with reference to the Facebook, and two different implementations are discussed for this purpose. First is the discussion of retrieving user data though the implementing a Cheri Facebook Application. And the second method of retrieving user information is by the implementation of the Cheri website. This module previously used the *Facebook connect method* but now uses the *Facebook graph* to achieve the purpose. The two implementations also helped test the issues with accessing user data while inside the facebook platform (Cheri facebook application) and while outside the facebook platform (Cheri's external site). The data collection process is explained by example.

Next in Section 6.2.2 the pre-processing of the data extracted from Facebook is discussed. This topic is already discussed in greater detail in chapter 5 here we only explained it by example in reference to the facebook data. A combination of limiter lexical filtering, spelling check and concept disambiguation is applied to filter the data. WordNet is used for identifying synonyms.

The next step of linking user interests with V&A via DBpedia and user profile formation is discussed in Section 6.2.3. The silent features of process are the reuse of known and widely used vocabularies and ontologies to add semantics to the user data and the use of LD principles for creating links. The mechanisms used are discussed in chapter 5 in greater detail.

This chapter identifies the various ways the recommendations can be presented to a user in Section 6.2.4. The Cheri system allows the user to visualise the results in *a product based view* and as *an active user location based view*. The product based view focuses on the attributes of the information that is to be presented to the user and the different ways of viewing it in the Cheri system e.g. the product based view allows the

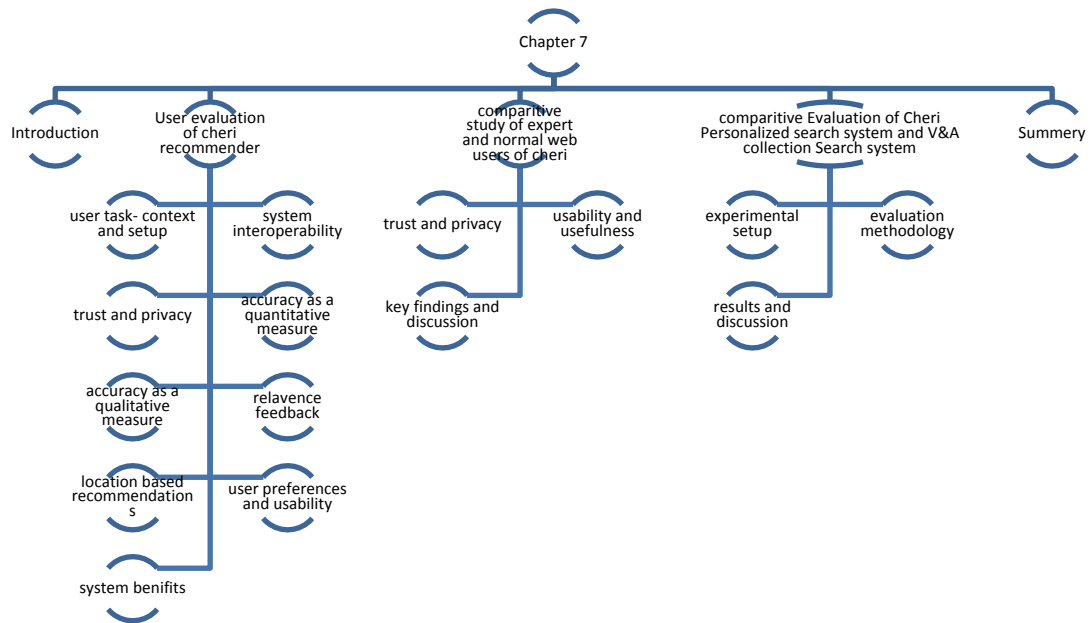
user to navigate through the suggested LOD and the V&A recommendations and explore it through various ontological concepts presented as filters. The product based view also presents another view that allows the user to explore the recommended artwork on a map, in this visualization the artwork is shown at their place of origin on the map and can be further explored by clicking on the image of the objet on the map. While the active user location based view takes into consideration the current location of the user while he/she is using the Cheri system and allows the user to explore artwork based on their current location and interests.

Section 6.2.5 gives the details of our RDF store ARC and a discussion to justify its suitability for the current system. Section 6.2.6 and 6.2.7 describe the implementation details of the Cheri recommender system and the Cheri search system respectively. The next chapter presents an evaluation of the two Cheri systems.



# Chapter 7

## Evaluation and Results



Chapter 7 topic hierarchy

### 7.1 Introduction

This chapter outlines the three evaluations designed to test the Cheri Art Recommender system and the Cheri Search System. The first evaluation comprises of experiments that were carried out to demonstrate the scrutability (user-acceptance and system accuracy) measure of the *Cheri Recommender* system, the dynamic and adaptive nature of the feedback mechanism based on the *Cheri* data model and the proposed idea of a *walking-museum*, that renders artwork according to its place of origin and user's current location and their interests. For each experiment, the methods and experimental results are described and discussed.



The second Evaluation is a comparative study outlining the analysis and observations made during the use of Cheri Recommender by a group of expert users (computer experts and researchers) and non-expert users (general Web user).

The third evaluation is an empirical study of the Cheri Search System in comparison to the V&A museum online search system (*search the collection*). It compares the Precision and Recall abilities of the two systems.

## 7.2 User Evaluation of *Cheri*

### 7.2.1 *User Tasks- Context and Setup*

**Evaluation design.** The *Cheri* system evaluation was conducted as a within-subject comparative user study with 21 participants (33% female and 67% male). These were a mixed group of people ranging from computer and communication science researchers to general Web users. It was necessary to have diversity in the data set of this study mainly because of the following reasons.

- Having only computer science students, who are expert fellow researchers and academics, certainly introduces bias into the data. But it was a voluntary choice, firstly because usability problems met by such qualified users can only be worse with less skilled computer users. And secondly to get a perspective of someone who is familiar with the technical aspects of the system was vitally important for this study.
- General Web user's opinion was important for this study because they are the intended end users of the system.

Ethical approval was obtained from the Ethical Committee of the School in order to get permission for the users to participate in the study. For each evaluation, participants were given the opportunity to use the system and register their thoughts about their experience with the system. To achieve this, three questionnaires (see appendix C) were provided to each participant that explained the objectives and their relevance to the study, assured the participant of anonymity, gave them the option of not participating in the study if they did not wish to and asked them to evaluate their

experiences while using the system. Any information asked which could be used to identify the participant, was kept separate from the experimental data.

**Participant's Profiles:** Individuals from 4 different institutions and 3 different research schools were identified as experts for the study and comprised 38% of the total participants; the remaining 62% of the participants were the general Web users.

The participants were chosen from personal research links and from people met at relevant events, such as the Annual Multi-disciplinary Research Showcase, University of Southampton. They were approached by email or in person and were free to accept or decline the request to participate in the evaluation.

**Setup and Procedure:** Precise instructions were provided to the participants that explained different stages of the user study and are summarised in this section. In order to complete the evaluation, the users followed the steps provided by the user task sheet (see appendix C). The main task was to visit the *Cheri* website and perform a set of steps to evaluate the performance of the system and to observe the adaptation of the *Cheri* system in response to those tasks and then answer a questionnaire comprising twenty-eight questions. Background information about user experience and perception of online museum systems was obtained through an initial questionnaire comprising eleven questions. A final questionnaire comprising five questions was provided to the participant to be filled at the end of the evaluation, which was designed to get the user's opinion about the security and privacy issues related to the *Cheri* system as well as ask about Web 2.0 applications in general. The evaluation steps are detailed hereafter.

Before the users were directed to the website, detailed instructions were given. The *Cheri* website was designed to guide the user interactively and step by step through the whole evaluation procedure.

**Step 1:** In order to make sure that the users understood the goals of the evaluation, a Participant Information Sheet (see appendix C) with an estimated reading time of three minutes was provided to the participants. It presented them with a summary of this research, the type of user data that was being collected, how this data would be used, the tasks required to be completed and information about the technical

requirements and their legal rights. In order to reduce the possibility of having outlier results, an opportunity was taken to verbally remind the participants that they should try to perform this evaluation as a normal routine browsing/searching task.

**Step 2:** The participants were then provided with the Pre-evaluation questionnaire, in which they were asked some general information such as their name, affiliation and familiarity with the internet and computers. This questionnaire also asked more specific questions like the user's inherent attitude towards museums, Web recommendations and online search habits.

**Step 3:** After the participant had completed the pre-questionnaire, they were provided with an Information Sheet regarding the tasks they were about to perform in order to test the system. The information page provided the list of tasks they should accomplish, the order in which they should perform the tasks, precise details on how to get the application running and a short explanation on how to give feedback while testing the system. On average, a set of two to four questions accompanied each task. These questions were targeted to capture the user experience while they used the system. The tasks and related questions were designed to gather information that would be useful to prove or disprove the hypothesis set in chapter 1 of this thesis. (For detailed task sheet and questionnaires see appendix C).

**Step 4:** After reading the Information sheet, the participants then practically performed the tasks as instructed in the information sheet. Users were then expected to execute the necessary actions to get the *Cheri* system up and running, according to the instructions. When the user opened the *Cheri* web-link it prompted the user to login to their Facebook account if the user was not already logged in so that the system could capture their explicitly mentioned interests from their Facebook profile. Once logged-in, the system prompts to the user what information the *Cheri* system will capture from their facebook profile and whether they are happy to share that information with the system. The interest information extracted includes the following:

- Explicitly mentioned interests in the user's facebook profile under the heading Arts and entertainment (Music, Books, TV, Movies) and Activities and Interests.
- Geo coordinates (user location at the time of use of the system)

This information is necessary for the *Cheri* system to personalise the recommendations of the artwork that it makes to the user. Once the user accepts the sharing request the system presents the personalised results to the user and the user is able to enter the *Cheri* art recommender and evaluate different functionalities of the recommender. Finally, at the end of the experience, the user is directed to the post evaluation questionnaire. These were aimed at summarizing the user's opinion and gave us their preferences on five selected pieces of the user's personal information (namely; location (country/city), hobbies and activities, interests, professional info and status updates), trust in online community and privacy.

**Questionnaires.** The entire user experiment was recorded in three parts: the pre questionnaire, the main questionnaire and the post questionnaire. The *pre questioner* addressed the users' familiarity with online CH recommender facilities and the use of the internet and personalization. The participants preferred interest seemed to be 'travel' for 33% of the cases, 'films/movies' and 'reading/books' for 29% and 'music' for 24% of them. Since the evaluation was based around the notion of exploring the user's interest in the light of online museums, the users were initially questioned about their experience with using handheld museum guide systems and searching cultural heritage related information through the internet. The study revealed that 38% of the users had never used a handheld tour guide system, 24% had ever visited a museum online and only 10% of the users have ever searched for cultural heritage related information online. The user trends for visiting online museums and searching for cultural heritage related information are shown in Figure 6.2.1. Of the 48% who have used a hand-held tour guide system most of the users tended to describe their experience as average to good as shown in Figure 7.2.2. On the other hand the study indicated that 82% of the users were interested in a personalised art recommender facility and said they would use it if it was made available. A few other questions aimed at determining the user's affinity for cultural heritage were asked, revealing that only 44% of the users showed an interest in physically visiting museums and exploring artwork there and 30% of those considered themselves art lovers (i.e. 13% of all subjects). These statistics indicate that an online art recommender will increase the viewing of the artefacts in the museum and will increase the visitors to the museum if not physical than virtually through an online facility. This will help

increase the awareness about the Art itself which is one of the main purposes of any museum.

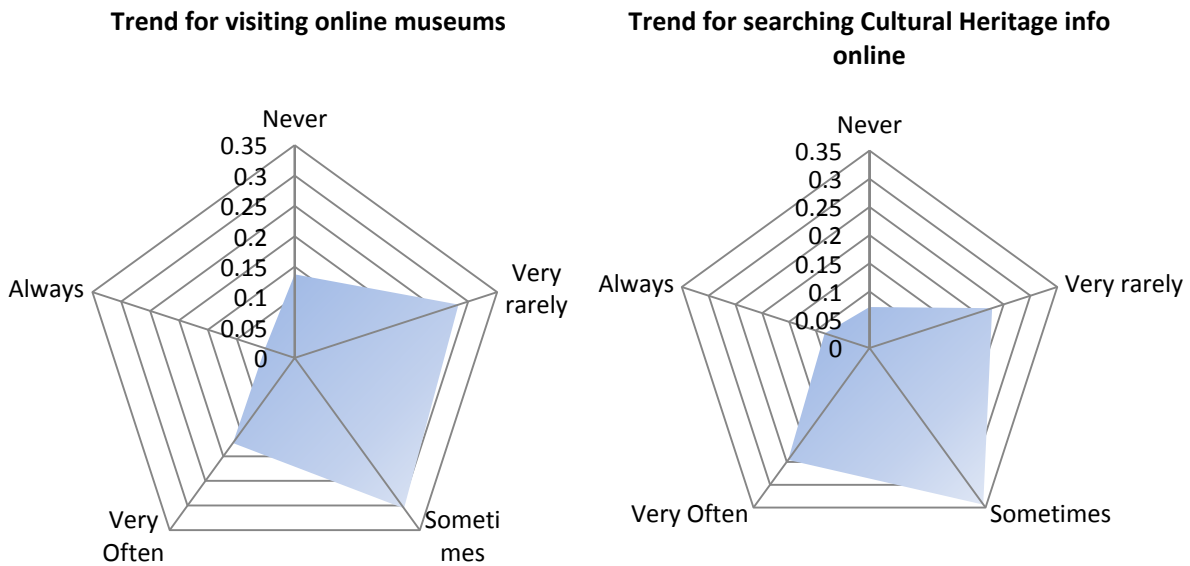


Figure 7.2.1: Measure of importance of privacy in social networks.

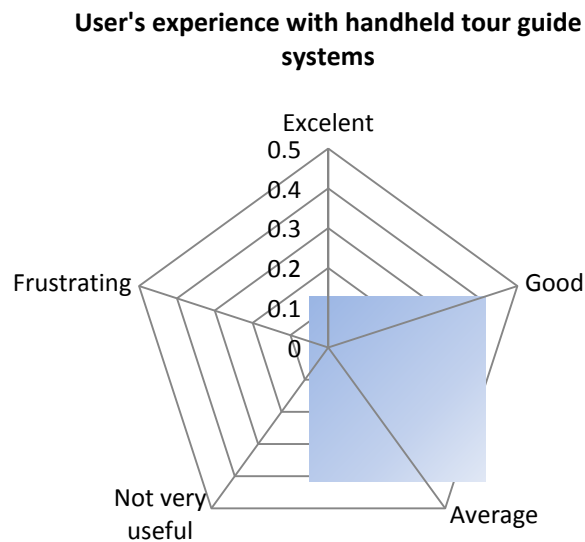


Figure 7.2.2: Measure of user experience with using handheld museum information systems.

The *post questionnaire* was designed to assess users' privacy and trust issues in a social Web environment, which was completed once the system had been tested. In the *main questionnaire*, questions were chosen according to a set of eight criteria that

we wanted to test (as proof of hypothesis in answering the issues detailed in chapter 1 and 3) in order to evaluate the *Cheri* CH recommender systems, namely:

1. System interoperability (covers hypothesis 2)
2. Trust and privacy (covers hypothesis 1)
3. Accuracy as a quantitative measure (covers hypothesis 3 and 5)
4. Accuracy as a qualitative measure (covers hypothesis 3 and 4)
5. Relevance feedback (System adaptation quality and feedback analysis )  
(covers hypothesis 5)
6. Usability (User preferences and usability) (covers hypothesis 1)
7. Geo results (Location based search) (covers our idea of walking museum)
8. System benefits

For each of these themes the following two types of parameters were studied:

1. Subjective variables
2. Objective variables

### 7.2.2 System interoperability

This section was designed to study the data capturing capability of the *Cheri* system from SNS.

The study focuses on Hypothesis No 1 of this research mentioned in Chapter 1 that states: *If the user is not asked to enter too much information about themselves and their interests to boot-start the recommendation process in a system, rather the system acquires it through users social networking activities, this can decrease the effort spent by the user, increase the ease of use of the system and help solve the cold start problem* (cold start problem is discussed in chapter 3 in greater detail). The research Questions (from Chapter 1) i.e.

- Whether it would be easy to capture the user interest data from a SNS and if so will the user find the process easy?
- Will the interest transfer from the users SNS be annotated and identified with the right concepts semantically? are also investigated here.

To investigate and answer these questions we considered system interoperability as part of the user evaluation of *Cheri* recommender.

As only the user can identify what they meant when they mentioned a certain interest in their SNS, we found it necessary to investigate the question of interoperability from a user point of view. Hence these questions were added in this user evaluation. A few of the variables here were subjective like the meaning of a particular user's interest and the liking and ease of the data transfer process. Others were Objective, e.g. number of interests correctly transferred from the SNS. The discussion below deals with the variables separately.

#### **Subjective parameters:**

The subjective question, asked for the interoperability between the *Cheri* recommender system and the SNS (in this case Facebook), was “if any problem was faced during the transfer of user interests’ data from Facebook to the *Cheri* recommender system”. The results were fairly satisfactory as we observed that about 95% of the participants faced no difficulty in the transfer of their interests from Facebook to the *Cheri* system. The results show a smooth transfer of user interest data indicating an ease of use. Upon enquiring the causes of problem faced by the 5% users who mentioned facing a problem during the data transfer we identified that the problems were not of a technical nature but in understanding of the representation of the data transferred. Some of the people who tested the system and were using Firefox faced issues in the display of the data transferred, which resulted in negative commentary. But once the process was explained and the system restarted the issue was resolved.

Table 7.2.1.1 shows the percentage statistics. The results for interoperability were highly satisfactory.

## Significance of results

Table 7.2.1.1: Subjective measures from system interoperability templates

Question Asked	Response Statistics
	<i>Percentage</i>
Did you face any problems during the transfer? [1=Yes, 2=No]	5% Yes 95% No

### Objective parameters:

The objective variables were aimed at obtaining impartial measures of what users saw as their transferred interest from facebook to the *Cheri* system and how efficient the *Cheri* system was in presenting the transferred user interest data from an end users perspective. Users were asked how many interests were added, or how much time it took. The templates that we collected gave precise indications on how many interests were correctly transferred from the Facebook profiles to the *Cheri* system.

Table 7.2.1.2 shows the results from the templates. The statistics were highly satisfactory. All users responded that this was the case.

### Significant results

Table 7.2.1.2: Objective measures from system interoperability templates

Question Asked	Response Statistics
	<i>Percentage</i>
Were your interests correctly transferred from your facebook profile to the Cheri system? [1=Yes, 2=No]	100% Yes 0% No
Were the suggested links displayed correctly? [1-Yes, 2-No]	100% Yes 0% No

The first attribute was about the semantics of the interests transferred. Users were asked if the interest transferred were correctly identified and visually presented. The results were 100%.



The second variable was the semantic link of the interest presented. The system displays to the user a link (URI) to the concept that it has identified as the meaning of a user interest. This is a crucial variable and must be verified as a single word may have multiple meanings or may relate to different concepts. The results here were satisfactory as 100% of the users found the concepts right. The third and important factor was the amount of time it took to transfer the interest and display them this was found to be on average 6sec for first time user and 1.5 seconds for returning user. Our system automatically recorded this time.

### 7.2.3 *Trust and privacy*

We found that people are quite reserved about privacy issues in general. But in practice they are more flexible about sharing their personal information on online through social networks. If a suitable incentive (e.g., meeting people who share the same interests) and a desirable gain (e.g., getting related information about their area of interest or ease of finding their desired knowledge) are offered, people show a tendency to share more of their personal information online. Similar are the observations we had from the results of the evaluation mentioned in this section.

We were also curious about what type of data are users most sensitive about as this is an important issue for a system that relies heavily upon user information to work. Following are the results and discussion of the subjective and objective variables related to the security and privacy related SNS issues for *Cheri* and *Cheri* like systems.

#### **Subjective parameters:**

The subjective questions asked related to the trust and privacy issue in an SNS environment included: How much privacy is an issue on the social networking sites? What is a user's befriending habits in an online community? What is a user's approach to sharing information online? Table 7.2.3.1 list the details of the statistical results obtained from some subjective variables considered to help answer the trust and privacy issues that are important considerations for designing *Cheri* like applications that rely on SNS.

## Significant results

Table 7.2.3.1: Subjective measures for SNS trust and privacy issues

Question Asked	Response Statistics
	Percentage
Are you comfortable in adding people you do not know in real life as your facebook friends?[1-Yes, 2- No]	29% Yes 71% No
Do you accept an 'add as your friend' request on facebook without knowing the person in real life?[1-Yes, 2- No]	38% Yes 62% No
How many of your facebook friend you do not know from your real life?	See Figure 7.2.2.2
<b>Moderately significant Results</b>	
How much is privacy in social networks important for you? Are you happy to share your information with other users of the social network? [1-Never, 2-Very Rarely, 3-Sometimes, 4-Very Often, 5.Always]	See Figure 7.2.2.1

The first factor established privacy as an important issue in SNS. Reinforcing the fact that due care is required while handling user interest information as we discussed earlier in chapter 2. The users were asked how comfortable they were in sharing information in a SNS environment. The results are moderately significant indicating an average inclination towards data sharing in an SNS environment as shown by the mean (and Standard deviation) of 3.09(1.33). 43% of the users indicated that they would share their information on a SNS site 'sometimes' while 28% of the people would rarely if ever share their information on SNS.

The second variable was how comfortable a SNS user is in adding people in their friend group if they are not an acquaintance from real life and what is the ratio of people whom a user just knows through online to those they have met in person. This is an important factor to study as it helps us understand the trust and privacy trends of SNS on a more personal level. The statistics show that a majority (70% to 60%) of people would not trust a person into their SNS circle if they don't know them in real life. On the other hand about 29 to 38% would try and make an acquaintance.

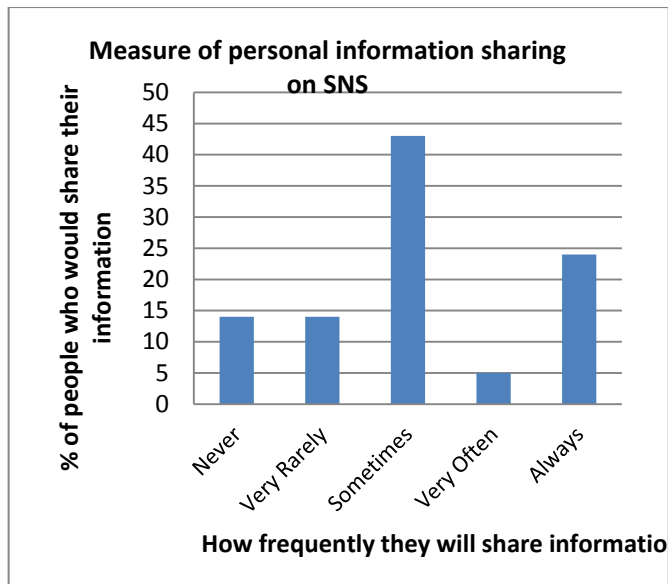


Figure 7.2.3.1: Measure of importance of privacy in social networks.

Figure 7.2.3.1 shows the measure of personal information sharing on a SNS. Most of the people agree that they would ‘sometimes’ share their personal information online with other members of their online social network. The online friend list of about 60%

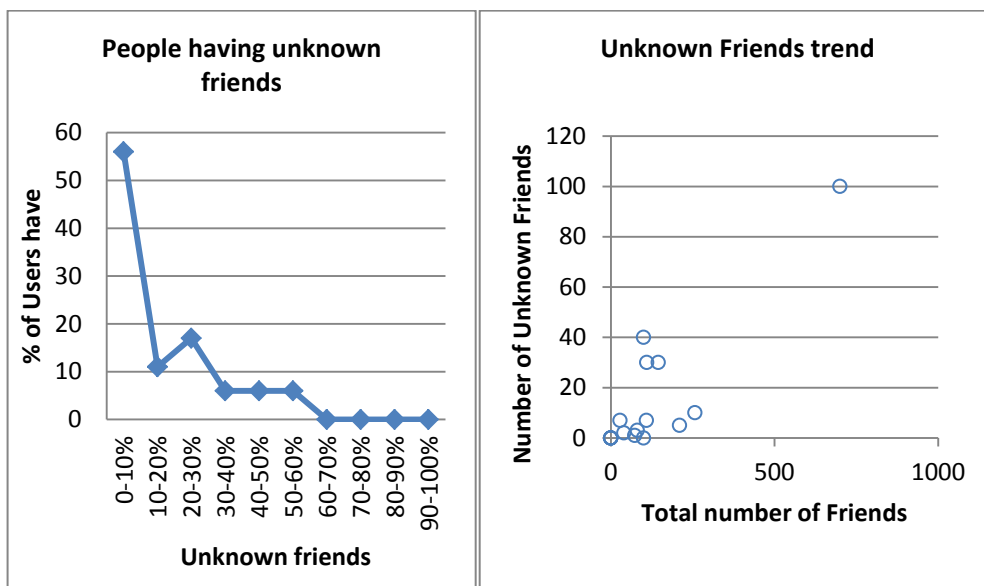


Figure 7.2.3.2: Percentage of people have unknown friends

of the SNS users may comprise of at least 0-10% unknown people and 7% of the users may have approximately 60% unknown connections in their SN circle as shown in Figure 7.2.3.2. The significance of these results may be compromised due to the selection of a relatively small data set. But the randomness in the selection of subjects does give us significant incite in the online behaviour of SNS users and the vast extremes found in the level of trust they put in the online community.

### **Objective parameters:**

The objective attribute considered here is significant to the study as this plays an important role in identifying the trust and privacy of information issues attached with the *Cheri* recommender system and the type of user information it requires in-order to

### **Significant results**

Table 7.2.3.2: Objective measures for SNS trust and privacy issues

Question Asked	Response Statistics
	Percentage
<p>What information will you share on Web with the following 3 categories of people tick as appropriate?</p> <p>1) With friends:</p> <p>Location (country/city), Interests, Hobbies &amp; Activities, Profession Info, Status Updates.</p> <p>2) With everyone:</p> <p>Location (country/city), Interests Hobbies &amp; Activities, Profession Info, Status Updates.</p> <p>3) With no one:</p> <p>Location (country/city), Interests Hobbies &amp; Activities, Profession Info, Status Updates.</p>	See Figure 7.2.2.4 and 7.2.2.3

provide personalised recommendations to its users. The user was presented with five categories of information that are related to them and asked which ones they feel they can share on an SNS system and with whom. Table 7.2.2.2 lists the categories of

information and people. The results of the analysis are described through Figure 7.2.3.3 and 7.2.3.4 below.

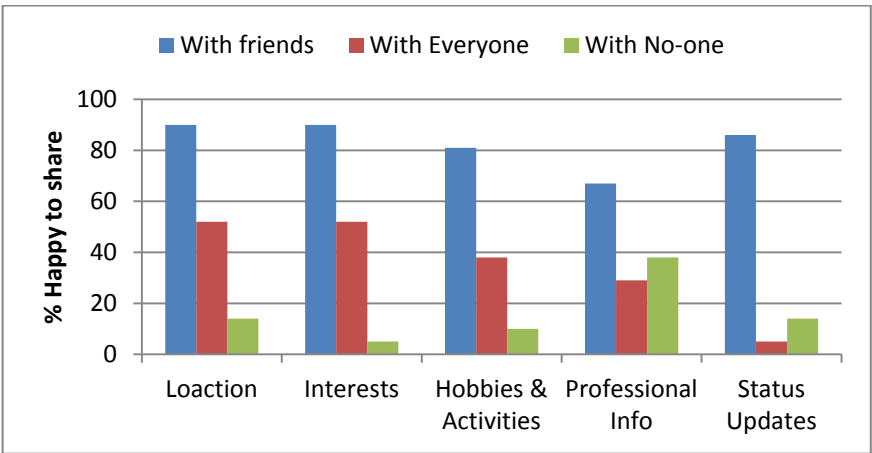


Figure 7.2.3.3: Type of information sharing

Figure 7.2.2.3 shows the information sharing trend of a selected set of SNS users.

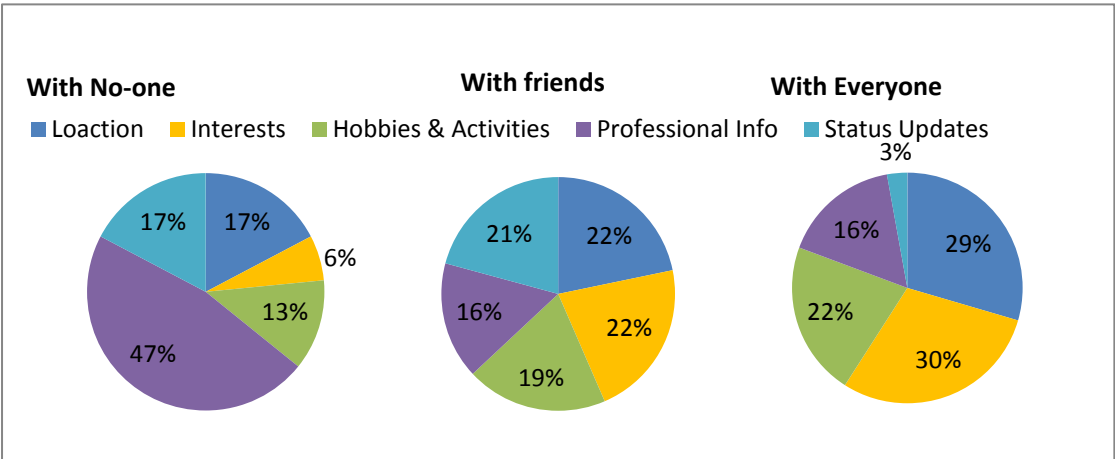


Figure 7.2.3.4: % Type of information sharing

Figure 7.2.2.4 shows most and least shared information on a SNS amongst different subset of SNS users.

Here, *Friends* are the trusted people on SNS whom the user knows well.

*Everyone* is the set of SNS users that are linked to a user’s profile. They may or may not be his/her trusted friends.

*No-one* is a set of people on the SNS that are not linked to the user's profile i.e. the information shared with this set of people would be open to the SNS, for anyone to view.

As seen from the pie-charts above most of the SNS users are equally likely to share their location, interests, hobbies, professional info and status updates with their friends online. Users are relatively hesitant to share their professional info with everyone on their list and even more hesitant to share their status updates. A more dramatic trend was seen in the last category as the results show professional info as the most private information that around 47% of the users were not willing to share with anyone on the Web.

The significance of this analysis is the fact that most of the people are willing to share their interests and location with everyone online. This discovery led us to the conclusion that sharing these pieces of information does not raise any serious privacy issues for majority of the SNS users. And most of the SNS users already trust their online community with these categories of information. Hence building a recommender system that relies on interest and location information of a SNS user will not pose any threat to the privacy and trust of information from a user's point of view.

#### 7.2.4 Accuracy as a quantitative measure

##### **Objective parameters:**

The objective variables for this section were the impartial measure of the interestingness from the user's point of view of the Artwork and related information from the LOD that was presented to the user as recommendations by the Cheri system and whether the users were interested in the recommendations. This section and section 7.2.5 supported Hypothesis 2 mentioned in Chapter 1 of this thesis which stated that '*Social web data can be used to gather up-to-date interest information about a user. The user's SNS interaction activities will better represent the user's ever changing interests.*' Users were asked how many interests were added, or how many of the recommended artworks they really liked. But from an accuracy point of view the most important factor was the number of concepts that the *Cheri* system was

unable to resolve i.e., semantically the *Cheri* system did not point to the correct concept and therefore the correct set of recommended items. The templates that we collected gave precise indications of how many interests the *Cheri* system was unable to resolve. The results are stated in Table 7.2.4.1 below. Table 7.2.4.1 clearly indicate that the user was able to resolve all concepts through the *Cheri* system.

### Significant results

Table 7.2.4.1: Objective measures for accuracy as a quantitative measure for *Cheri* system.

Question Asked	Response Statistics
	Percentage
How many concepts were you unable to resolve through the Cheri system?	0% Yes 100% No
Were the Web links relevant to your interest?[1-Yes,2-No]	90% Yes 10% No
Were the Web images relevant to your interest? [ 1-Yes,2-No]	90% Yes 10% No

### 7.2.5 User feedback mechanism and System adaptation quality analysis

The questions and tasks in this section of user evaluation were designed to demonstrate and evaluate the following key features of the *Cheri* system.

- The ability of the system to register the changes in the user interests.
- The ease of use of the feedback system.
- The effectiveness of the feedback mechanism.
- The measure of user satisfaction with the modified results.

The following scenario was designed to evaluate this part of the experiment.

**Scenario:** A person visits V&A (in person or online) and the system suggests him/her some artwork based on the terms from his/her interest profile e.g., painting, sports, archaeology. Our system suggests artefacts (e.g., pots with paintings of sports on them) from V&A if they exist. Of the suggested results the person likes a painting (or maybe the person adds it to his/her profile as his/her interests). It appears as facebook likes in his/her system.

The users are asked to see what features of the artefact they like contributed towards their inclination towards that object e.g., artist, time period, technique etc. The selected choices are used farther to suggest related results from the collection. The user is asked to evaluate the experience at the same time as they perform it. The user performs the following set of tasks: Task1: Try the *Cheri* system and see if you like what is suggested to you. Task 2: Register your feedback by the like option and choose the features of the artefact from the list of features (ontological concepts working as filters) that contributed in them liking that object. Task 3: See how the system has responded by modifying the results to your feedback and see if you are satisfied.

Based on the above mentioned set of tasks we designed the set of questions to be answered by the user. The categorical description of the questions asked and the research test results are given as follows. The accumulated results of this part of the experiment are given in table 7.2.5.1 and Table 7.2.5.2.

#### **Subjective parameters:**

The subjective variables were aimed at obtaining a qualitative measure of the user feedback collection method and system adaptation in response to the feedback. To measure these factors the users were asked about the ease of use and usefulness of the artwork property selection method which is used by the system as a tool for exploring the related artwork. The details of the questions asked and their statistical analysis is described below in Table 7.2.5.1

#### **Significant results**

Table 7.2.5.1: Subjective measures for system adaptation of *Cheri* through user feedback.

Question Asked	Response Statistics
	Percentage
Was the selection mechanism for modifying the search according to the artwork properties, easy to use? [ 1-Yes,2-No]	90% Yes 10% No
Did you find the feedback (through selection of properties) mechanism useful? [ 1-Yes,2-No]	88% Yes 12% No
Would you have preferred any other feedback mechanism? If Yes please state what other mechanism? [ 1-Yes,2-No]	24% Yes 71% No



Moderately significant Results	
Did you face any issues in using the Cheri feedback system? If Yes please explain? [ 1-Yes,2-No]	10% Yes 90% No

The first factor was designed to capture the user's views about the ease of use of the search and exploration mechanism, which comprises of a selectable set of properties about the recommended artwork, if a user likes a recommended object they are advised to select a single or a set of properties related to that artwork that they think were responsible for them liking the object. Once the user has made the selection the system automatically modifies the query incorporating the user's desired properties and presents the user with further results. The study showed that about 90% of the users found this exploration and query modification mechanism easy to use and 88% found the experience useful. Only 24% would have preferred some other mechanism. When asked about what method they would have preferred to explore or modify the results. They suggested they needed more choice in the properties by which the search can be modified (currently the system allows search modification through 4 properties related to art work). 10% of the users stated having issues with the feedback mechanism however when inquired the issues stated by the users were not of technical nature but what a user would prefer the Cheri feedback system to have in future e.g. it was suggested that '...it would be more interesting if a user can 'like' other objects from the suggested objects after the feedback by Cheri instead of only the attributes.' We aim to accommodate this suggestion in the future versions of Cheri. The results on the whole were satisfactory.

### **Objective parameters:**

The objective measures designed to evaluate the user feedback and system adaptation mechanism of the *Cheri* recommender focused on the evaluation of outcomes. The statistical details are given below in Table 7.2.5.2

## Significant results

Table 7.2.5.2: Objective measures for system adaptation of *Cheri* through user feedback.

Question Asked	Response Statistics
	Percentage
Were the results relevant to your feedback provided to the system through selection of properties? [ 1-Yes,2-No]	94% Yes 6% No

From the hands-on experience of the *Cheri* system 94% users recorded their satisfaction with the outcome of the search modification through feedback mechanism. Only 6% of the users felt that the modified results were not as relevant to the feedback they provided to the system as they would have liked. This discrepancy may have resulted due to the limitations of the collection.

### 7.2.6 Location and Place of Origin based recommendations

Apart from recommendations from V&A and interest based LD recommendations from LOD resources, the *Cheri* system also provides location (user's current location) and origin (place of origin of the artwork) based rendering of recommended artwork on map. To understand the usability of such facilities consider the following scenario.

**Scenario:** User is traveling through countryside or a city and he/she is curious about the history of that place. The user opens the *Cheri* recommender application on his/her mobile device. The application gets the IP coordinates of the person to identify the user location on the map and shows on the map the artefacts that were originally from (made at, discovered at and or are related to) this place and are now kept at V&A museum. The system then identifies the items that are most related to the users interest and if found such items are marked with red markers.

The applications of such a facility in a hand held device are vast both for the general user, a queries traveller or a working historian. Such a facility will help the user relate more personally to a new or otherwise not-that-interesting place. The application brings the museum to the user rather than the visitor to the museum. Thus making a

mobile device act as a *walking* museum, rather than a *walk-in* museum. This enables the user to explore the artefacts in the museum in a different way and place. The system shows results on Google map and also as a list of images with tags that are of interest to the user (from their interest profile).

Based on the above mentioned scenario we designed a set of questions to be answered by the user. The categorical description of the questions asked and the research hypothesis they helped to test are given as follows. The accumulated results of the evaluation are summarised in Table 7.2.7.1 and Table 7.2.7.2 below.

### Subjective parameters:

The subjective variables were aimed at obtaining a qualitative measure of the usability of the map based representation of the recommended artwork and the location based recommendation system. The results are stated in Table 7.2.7.1 below.

### Significant results

Table 7.2.7.1: Subjective measures for location based recommendations

Question Asked	Response Statistics
	Percentage
Did you face any issues in using the map based representation of the artwork? If Yes please explain? [1-Yes,2-No]	10% Yes 90% No
Did you face any issues in using the Cheri Geo based recommendation viewer? If Yes please explain? [1-Yes,2-No]	10% Yes 90% No

From the results one can see that the map based representation of the recommended artwork was to many-a-users' liking as only 10% of the users had issues with it. And while enquiring the causes we identified that some of the people who tested the system through Firefox faced issues in the map rendering, which resulted in negative commentary. But once the application was restarted the issue was resolved. Similar were the results obtained in the *current location* based recommendation testing where only 10% of the users were unsatisfied mostly because of the previously mentioned map rendering problem. Majority of the users about 90% had a satisfactory user experience.

### Objective parameters:

The objective variables were aimed at obtaining a quantitative measure of the accuracy of the location based recommendation and the origin based art representation on Google map. The users were asked a set of questions as they were evaluating these features in the *Cheri* recommender and the results thus obtained are summarised in Table 7.2.7.2 below.

### Significant results

Table 7.2.7.2: Objective measures for location based recommendations

Question Asked	Response Statistics	
		Percentage
Did the system register your current location i.e., is the map centred at your current location (e.g., Southampton)? [1-Yes,2-No]		89% Yes 11% No
Are the results presented as thumbnails over the map relevant to your current location? [1-Yes,2-No]		86% Yes 14% No
How many of the results out of total have red markers attached to them? [1-Yes,2-No]	See Figure 6.2.6.1	-

The first feature demonstrated if the system registered the current location of the user correctly. And the results show that for 89% of the cases the system correctly identified the user's current location. The second feature analysed if the results presented to the user were relevant to the current location or not. The user could test it by exploring the metadata and descriptions attached with the recommended art work. The results in this case were satisfactory as about 86% of the results were related to the identified current location. The next factor to be considered was how many of the results were tailored specifically to the interest as well as the location of the user. One should keep in mind that the interests of the users are varied and while filtering results with interests as well as location there is a huge possibility that many of the results satisfy one of the filtering criteria but not the other. Keeping this in mind the users were asked to look for the recommended art work on the map that had red coloured markers attached to them as those were the results that fulfilled both the filtering criteria. A location neutral search will only return the objects related to a user's interest while a location aware search will filter these results with a location based filter and show those that are related to that particular place and the user, with red

markers, indicates those results that are related to both the user's interest and his/her current location. The results from this exercise were interesting to observe as it gave us an insight in dealing with random uncontrolled user data and the possibility of finding something interesting in it. Figure 7.2.7.1 gives the trend of finding recommendations in V&A museum collection that fulfil both the criteria. Suggesting that most of the users had at least two such results that were related to both their current location and their interests.

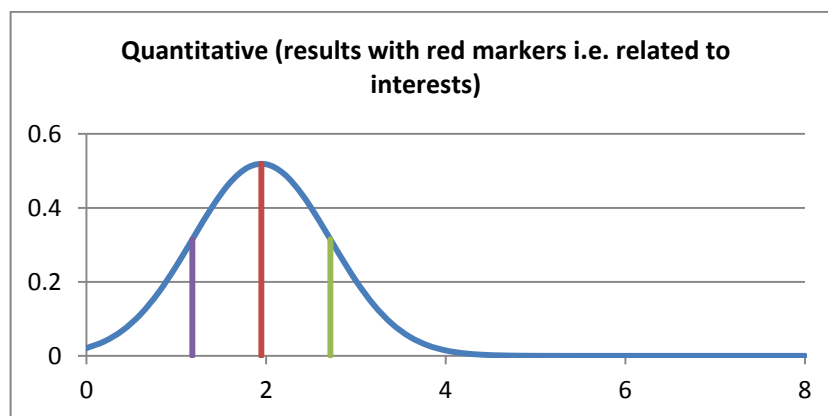


Figure 7.2.7.1: Interest and Location related Geo results from V&A collection

### 7.2.7 Usability (User preferences and usability)

Usability is an important factor to consider in any recommender system evaluation. To study the user experience of the *Cheri* users a set of questions were asked to record their satisfaction level and issues.

#### Subjective parameters:

The subjective variables were aimed at obtaining a qualitative measure of the user experience of the *Cheri* recommender and the evaluation process itself. Users were asked how easy or difficult they found the different tasks they were asked to perform while evaluating the *Cheri* system. And whether the different tasks were understandable. A summary of the results is given in Table 7.2.8 below.

## Significant results

Table 7.2.8: Subjective measures for user preference and usability of *Cheri* system.

Question Asked	Response Statistics
	Percentage
Did you understand the tasks that you were asked to perform? [1-Yes,2-No]	100% Yes 0% No
Did you face any difficulty in performing the task? If Yes State? [1-Yes,2-No]	10% Yes 90% No
Did you face any issues in using the map based representation of the artwork? If Yes please explain? [1-Yes,2-No]	10% Yes 90% No
Did you face any issues in using the Cheri Geo based recommendation viewer? If Yes please explain? [1-Yes,2-No]	10% Yes 90% No
Did you face any issues in using the Cheri feedback system? If Yes please explain? [1-Yes,2-No]	10% Yes 90% No
Did you face any problems during the transfer (of user interests from facebook to Cheri)? [1-Yes,2-No]	5% Yes 95% No

As the results in the table indicate the users found most of the features in the *Cheri* recommender easy to use (satisfaction levels between 90% to 95%).

### 7.2.8 System benefits

A few Open ended questions were also asked to record the users view about the applicability and benefits of the *Cheri* system. The suggestions are summarised in Table 7.2.9 below.

Table 7.2.9: Possible *Cheri* system benefits as suggested by users

Open ended Questions	Answers
How do you believe you can benefit from this system?	It attracts to know more about museum collection relevant to my interests
	We can know the history/information near us
	Get information about the collections that we might be interested in.
	Art is very far from my interest terms, but your suggested results show them in relevance to art, which introduces it to me.
	Relevant and quick search results.
	I can find interesting things relevant to my interests

	By finding links of your interests
	Helps to identify information more quickly
	Shows me my interests and activates nicely and in a good way
	I believe that the system will offer me information that will help me understand better about my interests.
	I believe it will be useful for students in the related research Shows me my interests and activates nicely and in a good way
39. What issues if any have you faced in using the <i>Cheri</i> system?	I am not familiar with it
	Adjusting my flow/understanding from one task to another
	Understanding the layout
	nothing

### 7.3 Comparing *Cheri Experience* of Expert versus Non-Expert Web Use

While analysing the results from user evaluation for the Cheri Recommender system we realised that the data collected showed an interesting insight amongst the responses of the expert and the non-expert users. How they adopt social networking technologies online and how they use the Web and Web applications. So a brief comparative evaluation of the results was done, the observations are discussed in this section as follows.

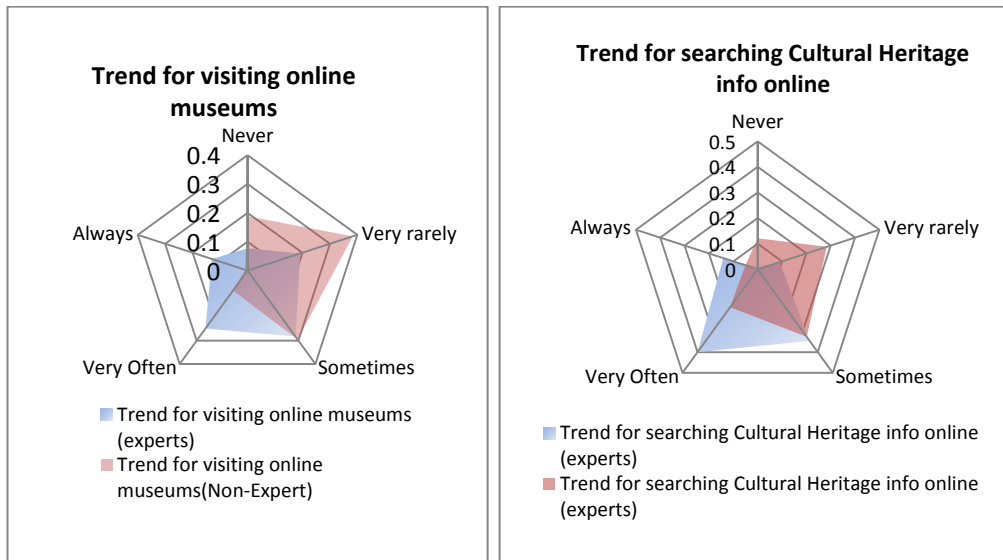


Figure 7.3.1 (a): Trends for visiting online museums and (b) Trends for searching CH information online

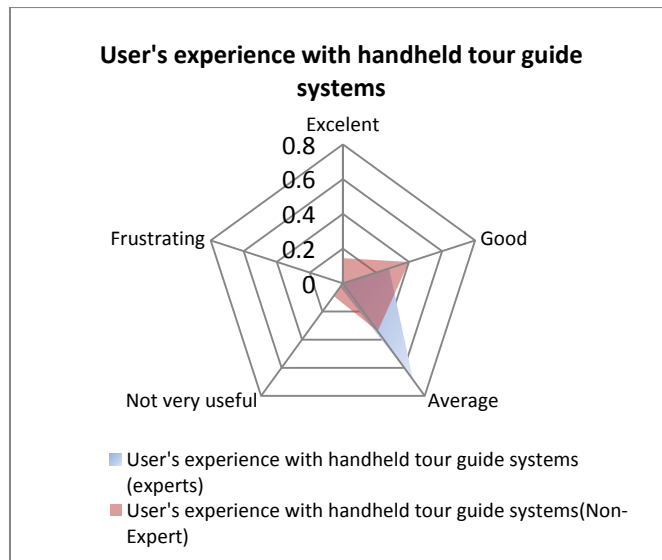


Figure 7.3.2: User's experience with handheld tour guide systems

As can be observed from the above three figures (7.3.1(a), 7.3.1(b) and 7.3.2) the non-expert user tend to not visit museum sites online or search for cultural heritage related information very often. This supports our argument for the need for Cheri like applications that unobtrusively introduce the user to selected artwork that is chosen keeping the user's interest in mind. An expert user however is more likely to navigate through such resources on their own. An interesting fact observed in Figure 7.3.2 was that the expert user is relatively less satisfied with the existing handheld tour guide systems than the non-expert user which may be due to the fact that an expert knowing the technology and hence the possibilities associated with it better.

### 7.3.1 Trust and privacy

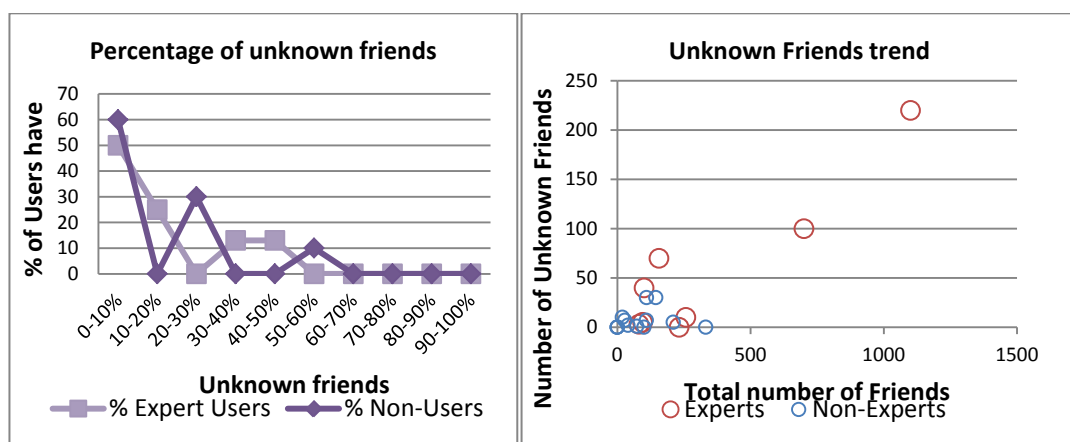


Figure 7.3.1.1(a) and (b): Trends for adding unknown people to one's friends list.



It was observed that the expert users generally had bigger networks of connections than non-expert users (as seen in Figure 7.3.1.1 (b)) however both sets of users had a remarkably high tendency for adding unknown people to their friends’ lists (as seen in Figure 7.3.1.1(a)). This implies that the information shared with a friend on SNS is quite likely to be seen by people outside the user’s trust circle. However the expert users are more sceptical of sharing their personal information on SNS than the non-expert (as seen in Figure 7.3.1.3 (a) - (b)). This may indicates a lack of awareness in general Web users regarding privacy related issues associated with SNS. This may also indicate that may be a general Web user is not that concerned about privacy (in practice) any way. However the researchers understand that the current user set for the evaluation is not big enough to make any generalised remark and so our comment holds true for the sub set of users in this study only.

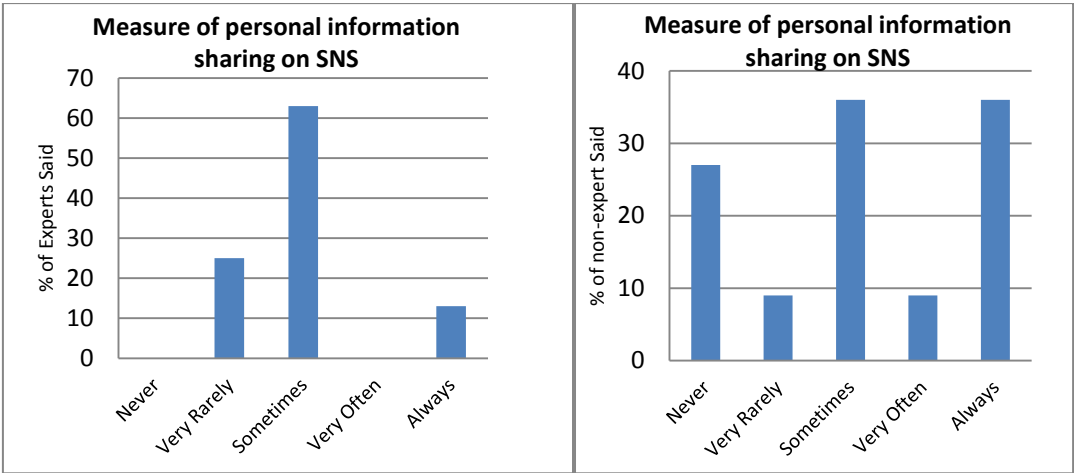


Figure 7.3.1.3 (a) and (b) Measure of personal information sharing on SNS

As observed from the trends demonstrated in Figure 7.3.1.5 (a) and (b) 100% of the experts are willing to share the type of information needed (in Cheri’s case location and interests) with their friends for personalised search and recommendation generation. While only 85% of the non-expert users would do so. However it is interesting to note that a bigger percentage of non-expert uses (55-63% approximately) will share the same information (location and interest) with everyone on SNS than the expert users (38-48%). These results are confusing yet important to mention here because they indicate the current state of trust and privacy related issues on SNS. And the fact that sometimes users are more comfortable sharing information

with people they are not acquainted with, how the user is sometimes not completely sure of what he/she wants or needs to share and what the consequences are (good or bad) while communicating on SNS.

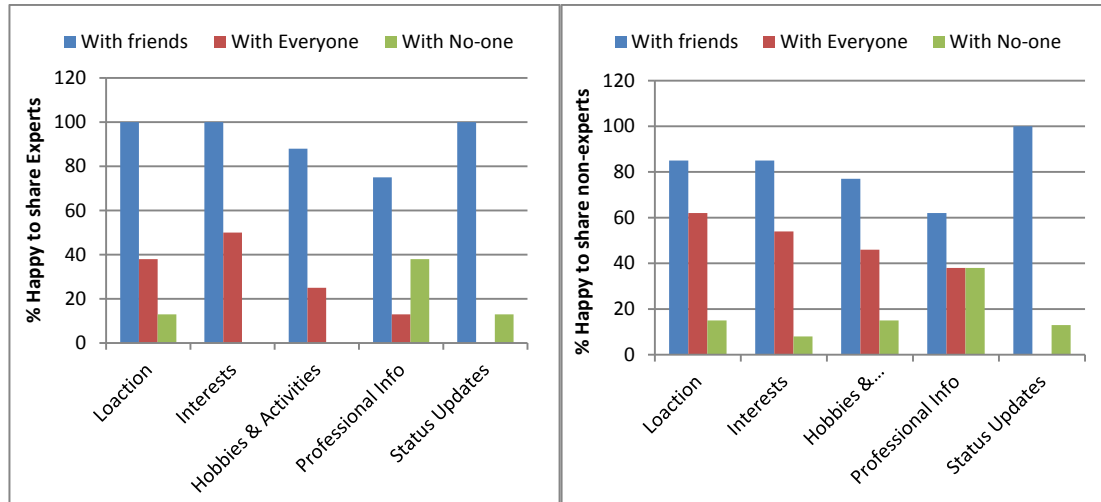


Figure 7.3.1.5 (a) and (b): measure of types of information shared on SNS

### 7.3.2 Usability and Usefulness

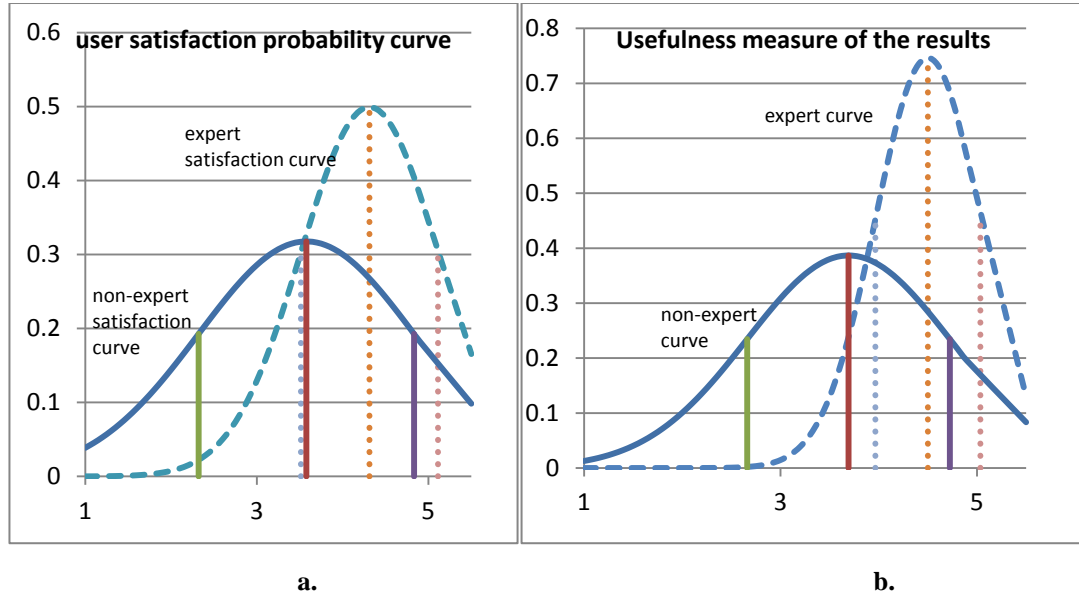


Figure 7.3.2.1 (a) User satisfaction distribution and (b) Usefulness measure for Web recommendations (y-axis: user frequency x-axis: (a) level of satisfaction (b) level of usefulness on the scale of 1 to 5)

Figure 7.3.2.1(a) shows that, in average, around 70% of non-expert people find user experience of the system to be at levels between 2.5 and 4.8 (which indicates average to high levels of satisfaction). While most of the expert users found the experience to be highly satisfactory. Similar trends were observed while measuring the usefulness of the results presented by the Cheri system to the users as shown in Figure 7.3.2.1(b).

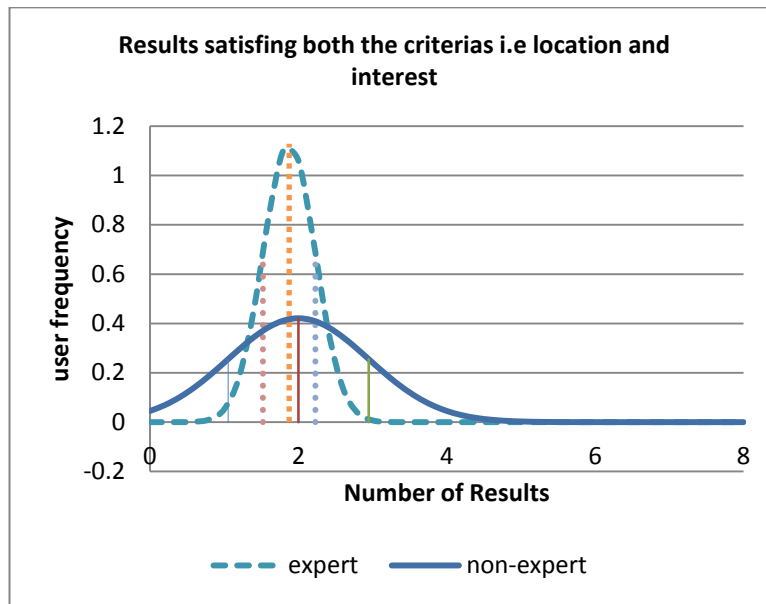


Figure 7.3.3.3 User satisfaction distribution

On average at least 25% of the results were found in the V&A data set that satisfied the criteria of matching user interest as well as had some relation with the user's current location. These results are important in situations when the user is using the Cheri system through a mobile device (e.g., smart phone). Although the expertise of a user had no direct link with receiving it is interesting to note how the distribution varies amongst the two sets of users.

## 7.4 Comparative Evaluation of Cheri Personalised Search System and V&A collection search System.

Performance can be investigated at several different levels, from processing (time and space efficiency), to search (effectiveness of results) and system (satisfaction of the users). Here we focus on evaluating retrieval effectiveness.

This is a comparative evaluation of the Cheri Personalised Search System and the V&A museum online collection search. The Purpose of the evaluation is to measure the Precision and Recall of the two search systems which are popular measures for evaluating retrieval effectiveness. The approach adopted is simple and to the point. A set of test queries are run through both the systems and the results obtained for each query are compared. A total of 21 queries were made to each system and Precision and Recall were calculated for both systems.

#### 7.4.1 *Experimental Setup*

The test query set was made up from a subset of 72 facebook interest terms that were shortlisted through the initial user evaluations of the *Cheri* recommender system and were found to be the most frequently occurring (see appendix A). The interest term was fed to both the Cheri search system and the V&A collection search API. The Cheri system applies the mechanism described in Figure 6.2.7.1 to generate recommendations. While the V&A museum London online search has its own search mechanism.

A discussion on the method used and the results gathered is as follows. The search systems were evaluated taking various cut-off points (1, 3 and 6 objects retrieved) pertaining to the *current user interests* for the estimation of Precision and normalized recall ratios for each pair of query and search system. The normalized recall ratios are taken to get a measure of retrieval effectiveness. The normalized recall ratio show if the search systems can display relevant documents in the top ranks of the retrieval output. If a search engine fails to retrieve any document for the search query the normalized recall value for the query will be zero.

The reason why an item is considered relevant also depends on the context. In the case of Cheri system the context to the search is user interest and location. Therefore, for a personalised cultural heritage (CH) query as done by Cheri the relevant items are those that are sorted according to the user's interest in addition to the relevance to the query term.

#### 7.4.2 *Evaluation Methodology*

Information retrieval (IR) research nowadays emphasises precision at the expense of recall (Tunkelang, 2009). Precision is a measure of the ability of a system to present only relevant items, i.e., the number of relevant items retrieved divided by the total number of items retrieved (Yates and Neto, 1999).

Precision and Recall were originally intended for set retrieval, but most current research assumes a ranked retrieval model, in which the search returns results in order of their estimated probability of relevance to a search query. However using precision at different cut-off points is helpful in estimating the distribution of relevant documents over their ranks (Bitirim, Tonta and Sever 2002). Other methods like mean average precision (MAP) and normalized discounted cumulative gain (NDCG) (Järvelin and Kekäläinen, 2005) are also used to reflect precision for the highest-ranked results. IR techniques like Kappa coefficient and gold standard or ground truth judgment of relevance are also widely used. But normalized recall was found to be more appropriate for evaluating highest-ranked result relevance measures in between V&A and the Cheri system. The Kappa coefficient is generally thought to be a more robust measure than simple percent relevance calculation since it takes into account the adjustment occurring by chance. However this very fact is regarded sometimes as a drawback of the kappa coefficient method (Strijbos, et al., 2006).

In gold standard or ground truth judgment of relevance, a document in the test collection is classified as either relevant or non-relevant, with respect to a user information need. However for the ground truth judgment the test document collection and list of information needs have to be of a reasonable size i.e., performance is calculated over fairly large test sets, as results are highly variable over different documents and information needs. As a rule of thumb, at least 50 information needs are considered to be a sufficient minimum. This was not the case for our evaluation.

Precision will always be an important performance measure, particularly for tasks like known-item search and navigational search. For more challenging information-seeking tasks, however, recall is at least as important as precision and it is critical that

the evaluation of information-seeking support systems take recall into account. Recall is the measure of the ability of a system to present all relevant items. i.e., number of relevant items retrieved divided by the number of relevant items in the collection.

Precision is closely related to the normalized recall which is denoted as  $R_{norm}$  (Yao, 1995). The normalized recall ratio shows whether search engines tend to display relevant documents in the top ranks of their retrieval outputs. If a search engine cannot retrieve any documents for a search query the normalized recall value for that query will be zero. The normalized recall is based on the optimization of expected search length (Cooper, 1998). In other words, it utilises the viewpoint that a retrieval output  $\Delta_1$  is better than another one  $\Delta_2$  if the user gets fewer non-relevant documents with  $\Delta_1$  than with  $\Delta_2$ . The normalized recall is calculated at three cut-off points (cut-off 1, cut-off 3 and cut-off 6) for each query per search system in order to be parallel with precision values. The  $R_{norm}$  is defined as:

$$R_{norm}(\Delta) = \frac{1}{2} \left[ 1 + \frac{R^+ - R^-}{R_{max}^+} \right] \quad \text{Formula 2}$$

Formula 2, proposed by (Bollmann, et al., 1986.), was used to calculate normalized recall values at various cut-off points. Here  $R^+$  is the number of document pairs where a relevant document is ranked higher than a non-relevant document;  $R^-$  is the number of document pairs where a non-relevant document is ranked higher than relevant one and  $R_{max}^+$  is the maximal number of  $R^+$ . Precision and normalized recall ratios were measured for each query on both search systems separately. Finally, these ratios are used to observe information retrieval effectiveness for finding V&A objects.

In particular, for tasks that involve exploration or progressive elaboration of the user's needs, a user's progress depends on understanding the breadth and organization of available content related to those needs. Techniques designed for interactive retrieval, particularly those that support iterative query refinement, rely on communicating to the user the properties of large sets of documents and thus benefit from a retrieval approach with a high degree of recall (Rao, et al., 1995). Meanwhile, information scientists could use information availability problems as realistic tests for user studies of exploratory search systems, or interactive retrieval approaches in general. The effectiveness of such systems would be measured in terms of the correctness of the

outcome (does the user correctly conclude whether the information of interest is available?); user confidence in the outcome, which admittedly may be hard to quantify; and efficiency i.e., the user's time or labour expenditure.

#### 7.4.3 *Experimental Results and Discussion*

This thesis conducts the study to see if the SNS data can be used to suggest artefacts that are related to the user interest, from online CH resources. So taking the query term set from SNS is necessary. The experiment uses the most frequently occurring words in the user interest profiles as a measure, with an intention that the queries generated from those interest terms will cover the interest a general SNS user will have. As for example, a list of most occurring search terms on Google is used to test a general search engine. In order to get realistic results only keywords were used as search terms as users do not tend to use phrases so often as observed in the most frequently occurring queries list of Wordtracker's "The Top 200 Long-Term Keyword Report", from 5th February 2008 and as observed in the most frequently occurring interest terms in our user data collection process (May 2011) figure 7.5 (see page 206).

Table 7.4.2.1: Query List

Query Number	Query	Query Number	Query
Q1	Sports	Q12	Film
Q2	cooking	Q13	Food
Q3	village	Q14	Football
Q4	books	Q15	Music
Q5	eating	Q16	Painting
Q6	parachuting	Q17	Reading
Q7	sleep	Q18	Social Web
Q8	Sports car	Q19	Swimming
Q9	flower	Q20	Technology
Q10	tennis	Q21	Tourism
Q11	teacher		

After each run of the query, the first 6 items retrieved were evaluated using binary human relevance judgment and with this every item was marked relevant or not relevant. A total of 252 items were evaluated by the same researcher and in order to have stable performance measurement of search systems, all the searches and

evaluations were performed in minimal non-distant time slots. While evaluating the retrieved items the following criteria were used:

- (1) Items that contain any explanation about the searched query were considered “relevant”;
- (2) In case of duplicated Items, the first item that was retrieved was considered in the evaluation process, whereas its duplicates were classified to be “non-relevant” (Bitirim, Tonta and Sever 2002); and
- (3) If, for some reason, a retrieved item became inaccessible, it was classified to be “non-relevant” ” (Bitirim, Tonta and Sever 2002).

Precision and normalized recall ratios were calculated at various cut-off points (first 1, 3 and 6 items retrieved) for each pair of query and search system.

Table 7.4.2.2: The number of relevant documents retrieved

Query Number	Cheri Search			V&A Search		
	Cut point 1 (Vis)	Cut point 3 (vis)	Cut point 6 (vis)	Cut point1 (dis)	Cut point 3 (vis)	Cut point 6 (vis)
Q1	1	2	4	0	1	2
Q2	1	2	3	0	0	0
Q3	1	3	4	1	3	6
Q4	1	3	6	1	3	6
Q5	1	3	4	0	0	1
Q6	1	3	6	1	2	2
Q7	1	3	6	1	3	4
Q8	1	3	2	1	2	4
Q9	1	3	5	1	2	5
Q10	1	2	5	1	2	5
Q11	1	3	5	1	3	4
Q12	1	3	3	1	3	5
Q13	1	3	5	0	0	0
Q14	1	3	6	1	3	5
Q15	1	3	4	0	2	5
Q16	1	3	6	1	3	6
Q17	0	0	3	0	0	2
Q18	1	3	1	0	0	0
Q19	1	2	3	1	2	5
Q20	0	1	1	1	2	2
Q21	1	3	6	0	1	5
Total	19	54	88	13	37	74
Avg (%)	90.4	85.7	69.8	61.9	58.7	58.7



Cheri	Precision	0.904	0.857	0.698	Average Precision= 0.819
	Average Recall	0.593	0.593	0.543	Average Recall= 0.576
V&A	Precision	0.619	0.587	0.587	Average Precision= 0.597
	Average Recall	0.406	0.406	0.456	Average Recall= 0.422

Recall is calculated by finding the recall for each query by the formula recall is equal to the correct result divided by the correct results plus the missing results. Then we calculate the average of the recalls for all the 21 queries.

The normalized recall ratios were calculated as follows for each pair of results for both the search systems at the three cut of points and are given in the following table.

Table 7.4.2.3: Normalized Recall Ratios for the two Search systems at 3 cut-off points

Query Number	$R_{norm}$ for Cut-off point 1		$R_{norm}$ for Cut-off point 3		$R_{norm}$ for Cut-off point 6	
	cheri	V&A	cheri	V&A	cheri	V&A
Q1	1	0	0.66	0.33	0.66	0.33
Q2	1	0	1	0.5	0.5	-0.5
Q3	0.5	0.5	0.5	0.5	0.6	0.8
Q4	0.5	0.5	0.5	0.5	0.75	0.75
Q5	1	0	1	0	0.7	0.1
Q6	0.5	0.5	0.6	0.4	0.87	0.37
Q7	0.5	0.5	0.5	0.5	0.8	0.6
Q8	0.5	0.5	0.6	0.33	0.8	0.6
Q9	0.5	0.5	0.6	0.33	0.7	0.7
Q10	0	1	0.5	0.5	0.78	0.36
Q11	0.5	0.5	0.5	0.5	0.72	0.61
Q12	0	1	0	1	0.5	0.75
Q13	1	0	1	0	0.9	-0.1
Q14	0.5	0.5	0.5	0.5	0.77	0.68
Q15	1	0	0.6	0.4	0.61	0.72
Q16	0.5	0.5	0.5	0.5	0.75	0.75
Q17	0	0	0.5	0.5	0.5	0.3
Q18	1	0	1	0	-1.5	-2.5
Q19	0	1	0.5	0.5	0.5	0.75
Q20	0	1	0.33	0.66	-0.16	0.17
Q21	1	0	0.75	0.25	0.77	0.68
Avg. $R_{norm}$ (cheri)	11.5/21=0.54		12.64/21=0.6		11.52/21=0.54	
Avg. $R_{norm}$ (V&A)	8/21=0.38		8.7/21=0.41		6.92/21=0.32	

As mentioned earlier Bollmann, et al. (1986) proposed the following generalised normalized recall:

$$R_{\text{norm}} = \frac{1}{2} \left( 1 + \frac{\text{the No. of agreeing pairs} - \text{the No. of contradictory pairs}}{\text{the maximum No. of agreeing pairs}} \right)$$

The number of zero retrievals (i.e., no items retrieved) or retrievals that contain no relevant items (i.e., the precision ratio is zero) can be used to evaluate the retrieval performance of search systems. The number of relevant items retrieved by each search engine in the first twenty one queries is shown in table 7.4.2.3.

Cheri have retrieved at least one relevant item for all queries but Victoria and Albert museum (vam) search has not retrieved any relevant document for three queries (2,13,18) i.e., 14% of the queries. For all queries, the cumulative percentages of non-relevant items Cheri and vam search have retrieved are approximately 18%, 40%, respectively. Thus, Cheri has retrieved approximately 50% more relevant items than vam search. Mean precision values of search engines in various cut-off points (1, 3 and 6 items retrieved) are shown in figure 7.4.3.1

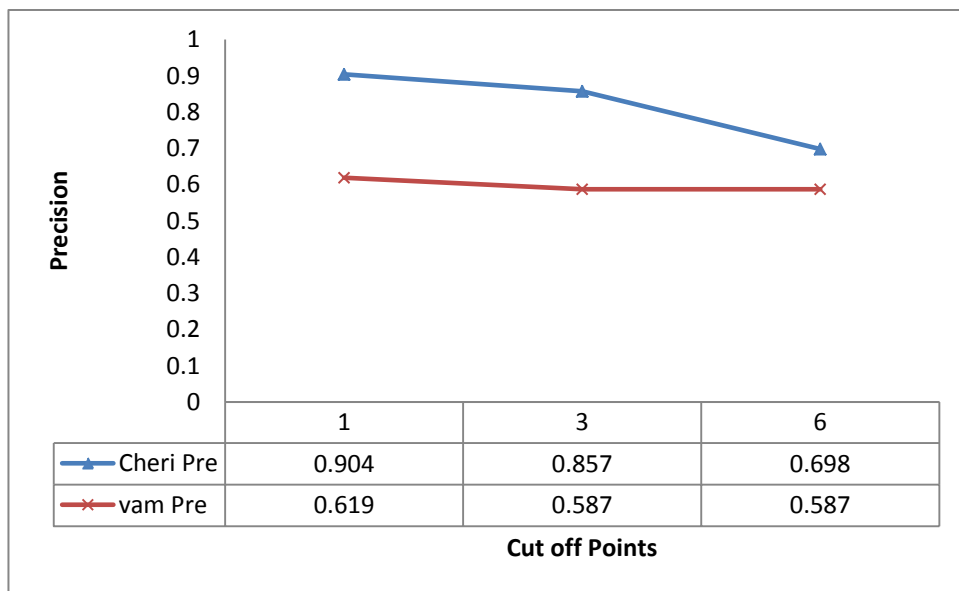


Figure 7.4.3.1: Mean precision ratios of Cheri and V&A search systems

When the cut-off point is increased, the precision ratios are decreased, which is a general trend observed while calculating precision at various cut-off points. Although Cheri's precision ratio show a greater decrease than vam with the increase in cut-off

point, Cheri still has the highest precision ratios on cut-off points 1, 3 and 6 (mean 81%), vam has (mean 59%) at the same cut of points.

Moreover, vam precision ratios on all cut-off points are lower than Cheri's precision ratios and the difference is approximately 19.6%. The mean precision ratio of Cheri is 81% on all cut-off points. Cheri has retrieved approximately 50% more relevant items than vam in cut-off points 1, 3 and 6. Thus despite the decline Cheri has still preserved its superiority in all cut-off points.

Mean normalized recall ratios of search engines in various cut-off points (for first 1, 3 and 6 items retrieved) are shown in figure 7.4.3.2

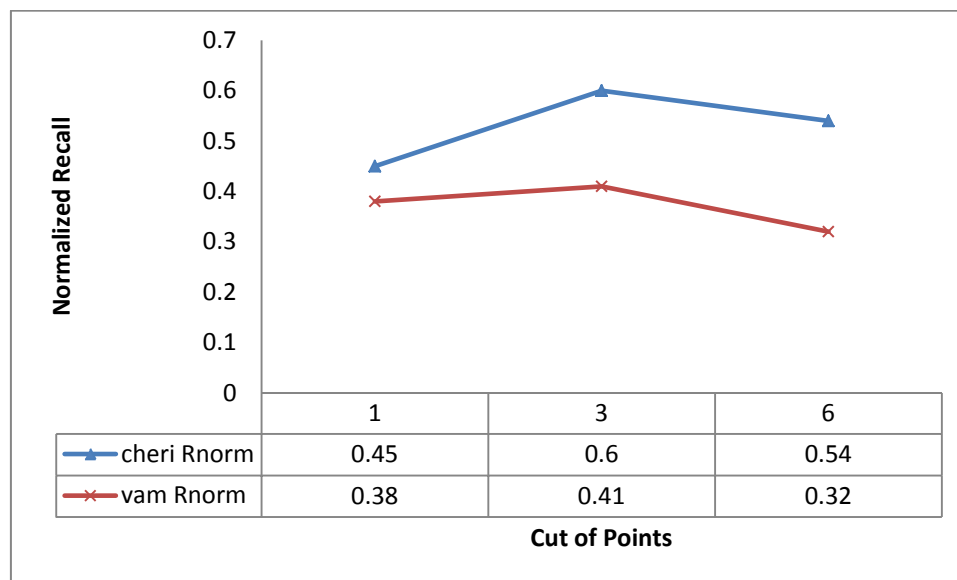


Figure 7.4.3.2: Mean normalized recall ratios of Cheri and V&A search systems

According to normalized recall measurements, Cheri has the highest performance at cut off point 3 (60%). The mean normalized recall ratio for Cheri is observed to be approximately 53% while for vam it is 37% with a difference of roughly 16%.

The major findings of this evaluation can be summarised as follows: Cheri retrieved more relevant museum items than vam's own search with the average of 50%. Mean precision ratios of the two ranged between 81% (Cheri) and 59% (vam). Cheri retrieved more relevant museum objects for all cut-off points. Mean normalized recall

ratio of Cheri is 53% and it means that, Cheri retrieved more relevant documents in the top ranks of the retrieval output when queries were run.

## 7.5 Summary

First evaluation was targeted to run as a proof of hypotheses discussed in chapter 3. It evaluated the Cheri recommender system on eight different aspects namely, interoperability, trust and privacy, accuracy (qualitative and quantitative), relevance feedback and system adaptation, location based recommendations, usability and system benefits. The interoperability aspect of the system yielded satisfactory results by indicating a 100% user satisfaction in interest transfer process between their SNS profile and the Cheri Recommender System. The process was found to be unobtrusive and easy with 95% of the users facing no problems during the process. Hence the first test for data extraction and interoperability passed the test.

The second aspect that was required to test the Cheri system was the privacy and trust issues related to Cheri or Cheri like systems/applications. Trust and Privacy are one of the most important aspects of any system working with user data, as the Cheri system feeds on SNS user data, this test was very crucial. We were particularly interested in identifying, what type of data the user is most sensitive about, as this is quite crucial for a system that relies heavily upon user information. The results regarding general data sharing are moderately significant indicating an average inclination towards data sharing in an SNS environment, as identified by the mean (and Standard deviation) of 3.09(1.33). It was also found that users are relatively hesitant to share their professional info with everyone on their friends list and even more hesitant to share their status updates. A more dramatic trend was seen regarding professional info as the most private information, around 47% of the users were not willing to share this information with anyone on the Web. The Overall significance of this analysis is the fact that most of the people are willing to share their *interests* and *location* with everyone online. This discovery led us to the conclusion that sharing these pieces of information does not raise any serious privacy issues for the majority of the SNS users. And most of the SNS users already trust their online community with these categories of information. Hence building a recommender system that relies on

interest and location information of a SNS user will not pose any threat to the privacy and trust of information from a user's point of view.

The third aspect, the Cheri System was tested for, is accuracy. The objectives for this section were aimed at obtaining impartial measures of what users saw as their recommendations and how good the *Cheri* system and its recommendations were from an end users perspective. From an accuracy point of view, the most important factor was the number of concepts that the *Cheri* system was unable to resolve i.e., semantically the *Cheri* system did not point to the correct concept and therefore, the correct set of recommended items. 10% of such cases were found during this evaluation. The user interest needed disambiguation in these cases, as it was not clear what they were actually referring to. However, the users were automatically given a set of possible choices by the system and were able to resolve all of such cases with one of the choices given. The user satisfaction probability was mostly between 3 and 5 (5 being the highest) with an average inclination of 4 and a standard deviation of 0.35 towards sigma right. The probability for usefulness of recommendations in user experience was found mostly between 3 and 5 (5 being the highest) with an average inclination of 4 and a standard deviation of 0.42 towards sigma right.

The fourth and fifth aspect of the Cheri Recommender evaluation was aimed at obtaining a qualitative measure of the user feedback collection and system adaptation in response to the feedback. This is an important aspect of a recommender system because it helps in refining results and automating query formulation to get user desired results. The results showed that about 90% of the users found this feedback collection and result modification mechanism easy to use and 88% found the experience useful. Only 24% would have preferred some other mechanism. When asked about what method they would have preferred to explore or modify the results. They suggested they needed more choice in the properties by which the search can be modified (currently the system allows search modification through 4 properties related to art work). From the hands-on experience of the *Cheri* recommender system 94% users recorded their satisfaction with the outcome of the search modification through feedback mechanism. The results on the whole were satisfactory.

The sixth feature of the evaluation tested the idea of a walking-museum. The implementation of the idea brings the museum to the visitor rather than the visitor to

the museum. By making a mobile-device act as a *walking* museum rather than a *walk-in* museum. This will enable the user to explore the artefacts in the museum in a different manner and location. A majority of the users (about 90%) had a satisfactory user experience.

Although usability was evaluated during the above mentioned six features, some specific questions related to usability and benefits of the system were also asked. The results indicated that the users found most of the features in the *Cheri* recommender easy to use (satisfaction levels between 90 to 95%).

Section 7.3 featured the second evaluation of the Cheri recommender system, where the users (evaluators) were divided into two distinct groups “experts” and “non-experts” and the way these two groups interacted with the Cheri system gave us a useful insight in the usability of Cheri as a real life recommender system rather than a prototype.

Section 7.4 presented the evaluation of the Cheri Personalised Search system which is essentially an extension of the Cheri Recommender System. This evaluation compared the performance of the Cheri Personalised Search system with the V&A (vam) museum’s *search the collection* search. The purpose was to find the improvements that Cheri provided on the original object search provided by the museum itself. The major finding of this evaluation was that Cheri retrieved an average of 50% more relevant museum items than vam’s own search. Mean precision ratios of the two ranged between 81% (Cheri) and 59% (vam). Cheri retrieved more relevant museum objects for all cut-off points in the query results. Mean normalized recall ratio of Cheri was 53%, which means that Cheri retrieved more relevant documents in the top ranks of the retrieval output, when queries were run.

Percent occurance of interest terms in facebook user data

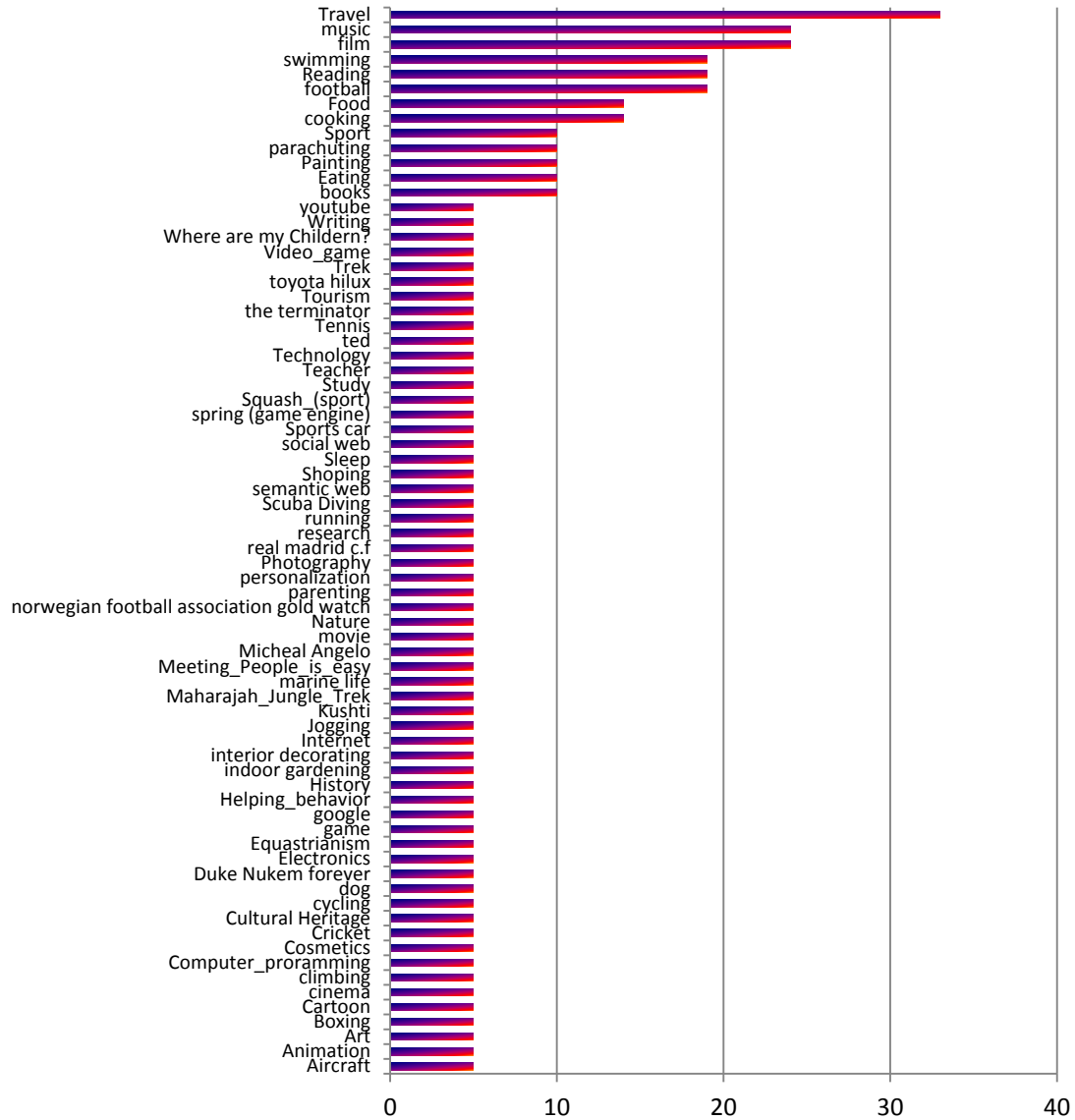


Figure 7.5. Percent occurrence of user interests in facebook data (used in Section 7.4)

## Chapter 8

# Concluding Remarks and Future Work

### 8.1 Summary of Research

**A framework has be developed to enable Cultural Heritage related Personalised Recommender Systems to consider Social Networking Data for dynamic user interest profile generation. Such a framework will contribute towards reducing the semantic gap between the cultural heritage expert domain knowledge and general Web user's interests. Moreover, as a consequence of the cross domain nature of the user profile, such a system will provide recommendations that are high quality, unexpected and geared solely towards satisfying user needs.**

It is argued in this thesis that the cold start problem is a common problem in personalised recommender systems and its root cause is lack of user interest information and or ways of capturing it. Also the problem of finding and updating user interest information unobtrusively and dynamically while relating them with appropriate concepts to suggest relevant information resources is still not solved. The thesis works around solving the above mentioned issues and overcoming the related sub issues as discussed in chapter 1 and chapter3. The solutions modelled, implemented and tested in a prototype recommender system called Cheri.

*What is Cheri?* The *Cheri* system is a user interest capturing, profile generating and art recommending system designed to make the Cultural Heritage domain more reachable to the general Web user. The interest profile generated through *Cheri* is mapped through LOD standards which make it reusable across the Web as well as machine readable. The interest profile is however layered with a mapping layer to



provide multi-domain knowledge. The system uses the interest profile to recommend artwork from the art collection of Victoria and Albert Museum, London that currently contains over a million records (V&A Search the collection, 2011), as well as open source information from DBpedia and the Web.

Below are the results from the proof of hypothesis that were proposed for this research and mentioned in Chapter no 1.

### **Hypothesis#1**

Hypothesis 1 stated that *‘When the user is not asked to enter too much information about themselves and their interests to boot-start the recommendation process in a system, rather the system acquires it through users social networking activities, this can decrease the effort spent by the user, increase the ease of use of the system and help solve the cold start problem’*. We saw in chapter 7 Section 7.2 in our user evaluation of the Cheri system that this hypothesis was supported by the result in table 7.2.1.1, Table 7.2.1.2 and Table 7.2.4.1 which indicate that; the users faced no problems in the transfer of interest information from there SNS profiles to the Cheri system, the interest terms were transferred correctly and the users were satisfied with the relevance of the interest terms transferred by the Cheri system based on the SNS profile, respectively.

### **Hypothesis#2**

Hypothesis 2 stated that *‘Social web data can be used to gather up-to-date interest information about a user. The user’s SNS interaction activities will better represent the user’s ever changing interests.’* We saw in chapter 7 from our observations mentioned in Table 7.2.4.1 that this hypothesis was supported by the results obtained and the users were satisfied with the relevance of the recommended results by the Cheri system based on the SNS interests terms. While the results from Section 7.2.5 mentioned in Table 7.2.5.1 and 7.2.5.2 on the ability of the system to register the changes in the user interests were also in support of the hypothesis.

### **Hypothesis#3**

Hypothesis 3 stated that *‘Ambiguity of SNS data can be clarified if their context is well defined and standard vocabularies and ontologies are applied to resolve this issue.’* We saw in chapter 5 how the Cheri System applies slandered ontologies like DBpedia and WordNet vocabulary to resolve ambiguities found in user’s SNS data, chapter 6 introduced the Cheri concept identification technique for resolving disambiguation in user data and in our experiment in Chapter 7 section 7.2 the technique is tested through the user evaluation and the results in Table 7.2.1.2 indicate that the Cheri system was able to disambiguate all such terms in the user SNS data hence this hypothesis was supported by the result.

### **Hypothesis#4**

Hypothesis 4 stated that *‘A generalised user interest-profiling system Based on users SNS data can serve as an interpretation of user’s interest and assist during recommendation or searching processes.’* We saw in chapter 5 that through the implementation of Cheri interest profile in FOAF ontology format the user profile is automatically made generalised and reusable. From our observations of the results of the user evaluation of Cheri system in Section 7.2 this hypothesis is farther supported.

### **Hypothesis#5**

Hypothesis 5 stated that *‘The profile thus generated will represent interests as concepts in a standard ontology and can serve as a useful resource for the recommender system in determining user’s interests and possible intentions while making recommendations, and in designing a mechanism for automated query formulation through the use of SN data.’* We saw in chapter 6 from the implementation of the Cheri Search and recommender system that such a system for automatic query formulation based on SN data is possible. This hypothesis is farther supported by the evaluations done in chapter 7 on the usability, precision and recall of the Cheri system.

The main novel contributions of this thesis are

- The work in this thesis (Cheri) not only introduces a new design space for SN (e.g. facebook) based personalised recommender systems, while working towards solving the cold start problem, but also explores and incorporates the lessons learnt from established techniques in the well-researched recommender systems, semantic Web and information retrieval domain as discussed in chapter 3.
- The significant difference between our work and that before it is the creation of a novel approach toward generating a dynamic and automated user interest profile from pre-existing user data in SN and populating it with related concepts from the open linked-data resources thus making it suitable for use across various context intensive information domains (such as cultural heritage) while addressing the cold start and related problems explained in section 3.2. This is the first project in our knowledge of this sort that helps personalise cultural heritage/museum search and recommendations using SN data.
- A novel filtering technique, that works by identifying the object types that a user will be most interested in viewing based on changing user interests and the cumulative weighting of the different objects presented to the user based on their interest. The filtering of the final ranked results based on the most related object types has been shown to bring variety to the results by ensuring that the top ranked results include items from different object types found in the V&A museum. The objects as well as the object types are selected based on the user interest and better represent the available knowledge in the recommender domain. For details see end of section 6.2.6.
- A novel concept of walking museum rather than a walk-in museum. The sixth feature of the evaluation tested the feasibility of this idea. The implementation of the idea brings the museum to the visitor rather than the visitor to the museum. By making a mobile-device act as a viewing medium for the art work based on the users current location and the place of origin of the artwork, this will enable the user to explore the artefacts in the museum in a different manner and location. The majority of the users (about 90%) had a satisfactory user experience of the feature.

## 8.2 Research Impact

### *Scientific Impact:*

- ***In Recommender system research:*** This research provides a fresh approach to providing a mechanism of avoiding the cold-start problem, which is a very common and major problem in the search and recommender systems domain. In addition the Cheri system places its self in a new class of hybrid recommender systems as evident from the analysis in table 2.6.1.2
- ***In Social Network Research:*** Two of the major problems with using social network data are; establishing credibility for the use of social data, in reasoning and query refinement tasks; and overcoming the sparse semantic structure of social data. This research establishes a case for Facebook as a representative of a broad online population of individuals, whose online personal networks reflect their real world connections. This makes it an ideal environment to study user interest dynamics and information contagion, which is a useful phenomenon considering one of the aims of this research is to introduce the general Web users to cultural heritage related information according to their interests in a seamlessly unobtrusive yet pervasive manner. Cheri in its implementation and evaluation has resolved the two identified issues successfully by making use of the interest terms from SNS to recommend artwork successfully to the user and by annotating the user data from SNS with DBpedia ontology and LOD to overcome the sparse semantic structure of SNS data.

### *Technical Impact:*

- As evident from the evaluation conducted in section 7.4 that compared the performance of the Cheri Personalised Search system with the V&A museum's (vam) *search the collection* search. Cheri provided significant improvement on the original object search provided by the museum itself. The major finding of this evaluation was that Cheri retrieved an average of 50% more relevant museum items than vam's own search. The mean precision ratios of the two ranged between 81% (Cheri) and 59% (vam). Cheri retrieved

more relevant museum objects for all cut-off points in the query results. Mean normalized recall ratio of Cheri was 53%, which means that Cheri retrieved more relevant documents in the top ranks of the retrieval output, when queries were run.

- The interoperability aspect of the system yielded satisfactory results by indicating a 100% user satisfaction in interest transfer process between their SNS profile and the Cheri Recommender System.
- From an accuracy point of view, the most important factor was the number of concepts that the *Cheri* system was unable to resolve i.e., semantically the *Cheri* system did not point to the correct concept and therefore, the correct set of recommended items. 10% of such cases were found during this evaluation. The user interest needed disambiguation in these cases, as it was not clear what they were actually referring to. However, the users were automatically given a set of possible choices by the system and were able to resolve all of such cases with one of the choices given.
- Feedback collection and system adaptation in response to the feedback is an important aspect of a recommender system because it helps in refining results and automating query formulation to get user desired results. The results showed that about 90% of the users found this feedback collection and result modification mechanism easy to use and 88% found the experience useful. From the hands-on experience of the *Cheri* recommender system 94% of users recorded their satisfaction with the outcome of the search modification through the feedback mechanism. The results on the whole were satisfactory.
- The results indicated that the users found most of the features in the *Cheri* recommender easy to use (satisfaction levels between 90 to 95%).

#### *Social Impact:*

- This work proposes and tests a way for opening the vast amount of structured data on Cultural Heritage to be exposed to the users of social networks, according to their taste and likings. One of the aims of this research is to

introduce the general Web users to cultural heritage related information according to their interests in a seamlessly unobtrusive yet pervasive manner. Through the user evaluations in chapter 7 it can be safely concluded that Cheri has achieved this goal to a reasonable extent, by successfully suggesting the SNS user Artwork related to their interest from V&A museum online and the LOD online.

#### *Economical Impact:*

- By using the social networking medium in the development and implementation of the Cheri system we inherently enabled the users to discuss and promote *Artwork* of their liking and to passively inspire friends. This can only help in promoting *Cultural heritage* and expanding the museum user-network which would mean more and more people virtually visiting the museums and making use of its resources. A positive economic impact can be speculated here for the cultural heritage industry.
- It is also more economical for the user to explore and discover artwork of their liking using a personalised cultural heritage recommender like Cheri. It is more time efficient and cost efficient especially for people who live abroad and cannot visit the museum itself or do not have time to do so. However the researcher does understand that nothing can replace the experience of viewing and spending the time admiring the actual artwork in person.

### 8.3 Contributions

This work contributed towards the following:

1. Our research has helped building a personalised search and recommendation system using strong semantics supported on standard semantic and social Web technologies, utilising the social Web as a context source. This generalised user interest profiling model helps the research system (*Cheri*) keep track of the changes in user's interests over time and incorporate these changes in the current search context accordingly and hence aid personalised

recommendations while avoiding some of the most well-known pitfalls in recommender systems as discussed in chapter 3.

2. Adaptive hypermedia and Adaptive Web research has been reasonably successful in exploring personalization in closed-corpus systems and to a lesser extent in open-corpus systems, but personalization on the Web is a complex phenomenon, extending beyond just content and encompasses many dimensions that need to be addressed consequently; for example social interaction, cultural preferences, and task and activities. This called for consideration of a multidimensional personalization model for the Web. The question arises as to how all these dimensions can be addressed in the same personalised experience without affecting or hindering the normal course of the search process. Our model provides a simple way to do so, i.e., by letting the users handle the diversity through the concepts they help us identify as their interests (encompassing divers topics like tasks and activities, cultural preferences and social interactions) in their SNS profiles and then using those concepts as a means of personalization while making recommendations.
3. Our portable interest model contributes towards a unified user experience across different sites, easy information access for service providing agents like recommender systems and end-user applications, increased recommender productivity due to less time required to search user related information (such as user interests), better planning of retrieval strategies and more accurate evaluation, better equipped exchange of user information across different platforms and above all meaningful personalization.
4. ***Introduce a fresh approach to solving the well-known cold start problem.***  
The cold start problem is the main problem to be solved in the context of the proposed framework. This is solved by ensuring that users are not assigned empty profiles upon registration, but rather carry with them the information that reflects their current interests across multiple domains. Of course, if users have not created any information prior to subscribing to the system (or have chosen to not disclose any) the problem persists. However, such behaviour

would somewhat defeat the point of seeking personalised recommendations. The user interest information is automatically gathered from the user's social networking account and is dynamically updated. And to avoid the initial and constant updating efforts required in making the user profile, linking the user interest model with the users SNs' profile is implemented as a solution.

5. ***Improve Findability and resolve the item similarity issue in recommender systems:*** The Cheri framework requires each resource to be mapped to a unique set of terms in the universal vocabulary (DBpedia). This provides a **mechanism for identifying interchangeable resources**. Such resources are expected to have identical descriptions using terms from the universal vocabulary (DBpedia) and can therefore be merged. This mechanism calculates the equivalence amongst items which is the basic solution for the item similarity issues, and increases the 'Findability' of previously hidden yet related information by querying all possible terms for the same concept (synonyms) aiding in **new knowledge discovery**.
6. ***Our novel interest filtering technique*** proposed and implemented in the Cheri system (besides solving the problem of shifts and temporal cycles of user interests) also has shown good results in helping to elevate the 'similarity of item' problem common in recommender systems, by ensuring that the recommended results are always from a set of highly weighted resources across a set of resource types (best representing the user's current interest) rather than from a single type of resource. In our proposed filtering model the query results are presented in order of relevance, but to avoid the most *similar items are not always good recommendation* phenomenon, our novel approach of filtering the results thus obtained with the current top five resource types (from the domain) that the system has calculated to be most related to the user current interest profile, has shown good results. This has been shown to bring variety into the results without losing relevance to the user, as can be seen from the results of evaluation in section 7.4.3. In addition the automatic upgrading of the interest profile each time a user logs a new interest in their SN ensures a dynamic interest profile that forms the core for the recommender



system making sure that the current interest is always considered while handling a user search or query string.

7. The Cheri framework successfully presents a *novel solution for a potential biasing effect in recommender systems* by shifting the emphasis to satisfying user needs. By introducing a standalone user preference/interest calculation and updating mechanism independent of the end data resources, it becomes harder to spuriously insert an arbitrary recommendation. Moreover, to influence the system to recommend the said resource over others, one would also have to obtain control over the universal representations of resources and the semantic connections between their descriptive terms in the universal vocabulary. Furthermore, since the SN data are simply seen as platforms indicating the preferences of their members, there is no guarantee of what objects will be selected for a user, on the bases of extracted SN data, as a recommended resource.
8. On its launch, neither Facebook nor the publishers (its partners) did any mark-up on their pages. At the time none of the entity pages on Facebook.com had Open Graph mark-up and thus Facebook's own pages remain closed. Ironically, this might not be because the company does not want to mark-up the pages, but it might be because it cannot until it figures out what is actually on the page. This is what semantic technologies have been working on over the past several years. In this thesis we *introduced a feasible way of marking up user data on the Facebook graph via a universal vocabulary (DBpedia)*, though not unique to semantic Web research, it would be the first time to suggest it as a solution for a big SN graph.
9. In the issues with recommendations made independently of context it is realised that the object/resource attributes alone are not adequate for representing the context of a recommendation. The framework offers a solution for automatically determining which aspects of a user interest profile are relevant to the context of a particular query. It achieves this by providing a *novel search tool* which overlays the user current interest rating with the

context of the user query to produce results explicitly selected to reflect a particular context. The search results are filtered through a user interest matrix. Any resource type that framework finds related (through semantic annotation and ranking) to a user interest, (that is, the user has implicitly expressed interest in it as part of their profile, regardless of their origin) is considered. By adopting this mechanism, the effects of problems associated with the inadequacy of user profiles to represent a wide range of user interests are expected to be less severe.

## 8.4 Future Work

### 8.4.1 *Deeper Semantics equals better exploration*

**Work in progress:** To explore a highly contextualised domain such as that of cultural heritage in greater detail a much richer context model is needed. To achieve this we intend to map the DBpedia concept ontology (our current universal vocabulary) to the CIDOC CRM model which is a standard model for CH information interchange. This mapping will provide us with a better means to exploit the vast amount of semantic information locked inside different CH bodies around the Web; thus making it rich and more widely applicable. The following diagram Figure 8.4.1. (b) Shows an entity relationship model of some of the concepts from the CIDOC CRM model. The figure describes the entity *Man Made Objects* E22 along with some other entities (classes) and the relationship that they share. As you can see the *Type* class (E55) is extended by a subclass from the AAT (Arts and Architecture Thesaurus) that is a well know Getty classification in the cultural heritage domain. The extension makes the CRM more expressive and thus more suited for our experiment. The subclass *fiction* is actually a sub class of the class *genre* in AAT.

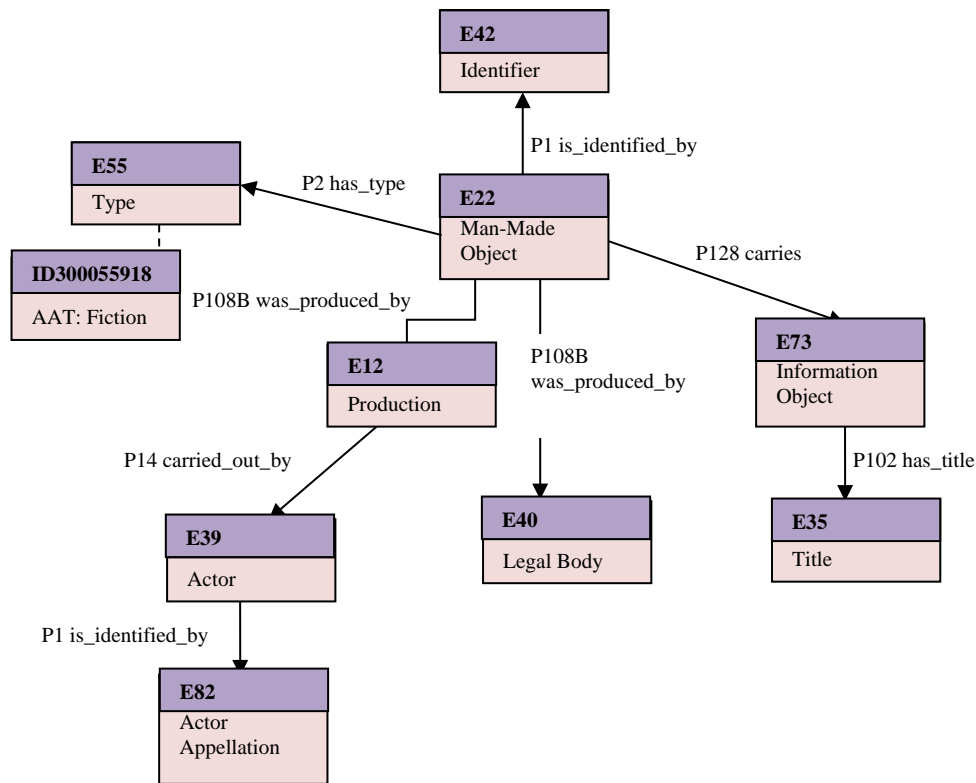


Figure 8.4.1 (b): Example CIDOC CRM Representation of the matching properties

Furthermore Figure 8.4.1. (c) describes the possibility of mapping between the above mentioned CRM concepts to the DBpedia concepts from our example of “The Lord of the Rings”. Thus, making a transition from an upper-level more generalised ontology to a domain level precise ontology. The mapping between the two ontologies lies in the following relations.

dbpprop:genre owl:same\_as E22-P2-E55

dbpprop:author owl:same\_as E22-P108B-E12- -E39-P1-E82

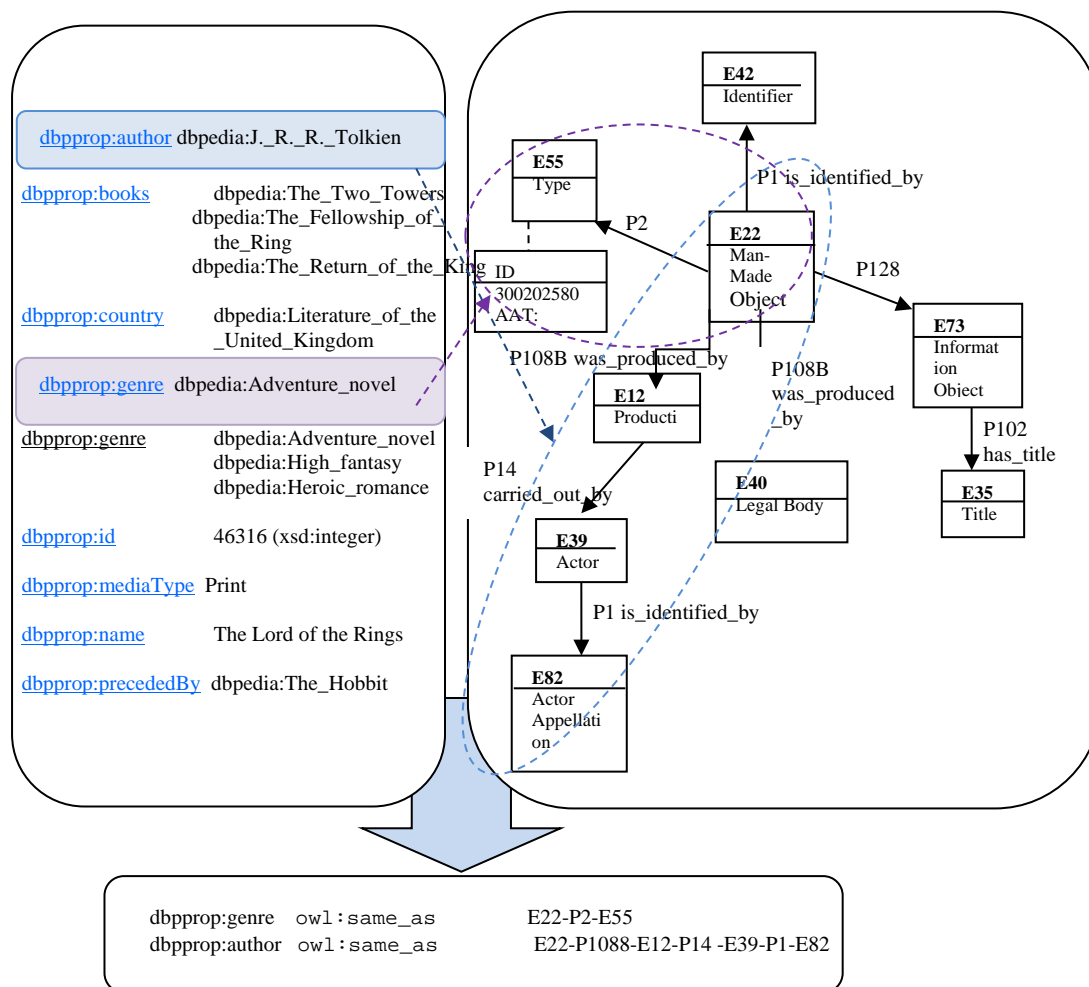


Figure 8.4.1. (c): Example Mapping Process.

We omit the details of the infrastructure for mapping between the CIDOC CRM conceptual model, vocabularies used (e.g., AAT) and gazetteers for the current discussion, but give a simplified example to prove the point. The example shows how CRM, AAT and DBpedia are incorporated together to relate a book from the user interests to its author. This complex representation will help exploit the relations amongst the user's social Web data and assist in improved reasoning over it, for better recommendations.

This example gives us a useful insight into the possibilities that lie in exploiting the hidden links in user generated public data from the social Web and highly contextualized cultural heritage data online. Our research aims to bring forward some

of the many possibilities that lie in exploiting this enormous and ever growing resource of data for useful yet interesting Web implementations.

#### *8.4.2 Scalability and Seamless Integration*

In future we intend to integrate the Cheri system with other data sources like BBC graph API and music API to see how the system acts with other domains and open-link resources. We also plan on extending our experimentation with other cultural heritage repositories by incorporating British Museum online. The V&A's collection also includes the National Art Library and the Archive of Art and Design, which are catalogued on the library database also available through the Museum website (<http://catalogue.nal.vam.ac.uk>). We eventually intend to incorporate it in the system as well.

The research has provided an initial analysis and design to the wider vision of next generation personalised CH recommender systems. While this might be easier to study under the context of ecommerce and business where the data belongs to a specific domain, the variation of data in personal profiles in SNS makes it more challenging in the personal domain.

#### *8.4.3 Data Extraction: extending the user interest gathering domain*

Some work that we have already done in expanding the user data collection process beyond Facebook is given in Appendix B. We intend to incorporate the method in future Cheri experimentation.

We believe that by linking all the different social identities of an individual over the Web and by unleashing the vast amount of contextual information enclosed in them, a richer and dynamic model of user interests can be achieved. That can serve as a rich context to further assist adaptive and user oriented applications and search processes. Unified profiling and tag data portability efforts are a way forward in this direction.

#### *8.4.4 Use of Extended Filters*

One of the observations with the current Cheri system user evaluation was the need for the provision of more filters for data exploration. That was a feature explicitly

demanded by some of the users indicating that it will improve the user experience and aid exploration. We intend to answer this issue in the next version of the Cheri system.

#### 8.4.5 *Improving User Control and Reasoning*

The system presently generates automated recommendations from the V&A museum for the users and suggests interesting pictures from flickr and information from DBpedia and is improved through soft feedback (like), ranking and filtering (using ontology concepts). In future we are planning to expand this by providing a greater level of inference over the V&A data through mapping of the user profile with Cidoc CRM and the use of OWL DL and pOWL. The work on this is currently in progress.

#### 8.4.6 Cheri Mobile Application and concept of walking museum:

The V&A data visualiser presents two map based representations of the results. These options were provided because we intend to introduce the *Cheri* system as a mobile based application in future. And the map based rendering of the artefacts will help us provide the facilities of using *Cheri* as a walking museum as well as a means of finding the cultural heritage of a new place while visiting it. The product based visualization is provided under the *Place of Origin* tab in the *Cheri* system as shown in Figure 6.2.4.3. This option shows each selected artefact at its place of origin, i.e. the place it was made or first discovered. This is an interesting option for a general user and a useful one for a working archaeologist or a historian. An *active user* is a user who is currently logged-in to use the system. The active user location based recommendations refer to the set of recommendations that are based on the current location of the user in addition to the active user interests. The recommendations are presented under the *near you* tab in the *Cheri* system and represent the artwork from the V&A museum that has originated from or is related to the users current location.

These implementations and their evaluation success provided us with the reassurance we needed regarding the usefulness of Cheri as a mobile application. Although one can use Cheri on a laptop, releasing a version for more portable devices like mobile phones will be our next target.



# References

- Abowd, D. A., Atkeson, C. G., Hong, J., Long, S. and Pinkerton, M., 1996. Cyperguide: A Mobile Context-Aware Tour Guide. *Wireless Networks*, 3(5): pages 421–433, 1996.
- Adamic, L. A., Buyukkokten, O. and Adar, E., 2003. “A social network caught in the Web,” *First Monday*, vol. 8, no. 6, June 2003.
- Agarwal, A., 2007. To Download Contacts from Facebook To Outlook Address Book. *Digital Inspiration*. October 15, 2007.
- Airio, E., Järvelin, K., Saatsi, P., Kekäläinen, J. and Suomela, S., 2004. CIRI - an ontology-based query interface for text retrieval. *Web Intelligence: Proceedings of the 11th Finnish Artificial Intelligence Conference*, Hyvönen, E., Kauppinen, T., Salminen, M., Viljanen, K. and Ala-Siuru, P., editors, September 2004.
- Altinel, M., and Franklin, M. Efficient Filtering of XML Documents for Selective Dissemination of Information. In *Proceedings of VLDB*, pages 53-64, Cairo, Egypt, September 2000.
- Angeletou, S., Sabou, M. and Motta, E., 2008. Semantically enriching folksonomies with FLOR. In *Proc of the 5th ESWC. workshop: Collective Intelligence & the Semantic Web*, Tenerife, Spain, 2008.
- ARC2, 2011. Easy RDF and SPARQL for LAMP systems|ARC RDF classes for PHP.[online] Available at:<<https://github.com/semsol/arc2/wiki> >[Accessed 15 May 2011].
- Argerich, L., 2002. Parsing RDF documents using PHP- An introduction to the RDF XML syntax and how to parse RDF documents using the PHP. version of Repat. 2002.



- Aspan, M., 2008. How Sticky Is Membership on Facebook? Just Try Breaking Free. New York Times, February 11, 2008.
- Athanasis, N., Christophides, V. and Kotzinos, D., 2004. Generating on the fly queries for the semantic Web: The ICS-FORTH graphical RQL interface (GRQL). Proceedings of the Third International Semantic Web Conference, Nov 2004, pages 486–501.
- Athanasis, N., Christophides, V. and Kotzinos, D., 2004. Generating on the fly queries for the semantic Web: The ICS-FORTH graphical RQL interface (GRQL). Proceedings of the Third International Semantic Web Conference, Nov 2004, pages 486–501.
- Auer, S., Bizer, C., Lehmann, J., Kobilarov, G., Cyganiak, R. and Ives, Z., 2007, Dbpedia: A nucleus for a Web of open data. Proceedings of ISWC07, 2007
- Auer, S., Dietzold, S., Lehmann, J., Hellmann, S. and Aumüller, D., 2009. Triplify – Light-Weight Linked Data Publication from Relational Databases. Proceedings of the 18th World Wide Web Conference (WWW2009).
- Backstrom, L., Huttenlocher, D., Kleinberg, J. and Lan, X., 2006. Group formation in large social networks: Membership, growth and evolution. In Proceedings of 12th International Conference on Knowledge Discovery in Data Mining, pages 44–54, New York, NY, USA, 2006.
- Baker, G., 2008. Free software vs. software-as-a-service: Is the GPL too weak for the Web?. Free Software Magazine. May 27, 2008.
- Baldoni, M., Baroglio, C. and Henze, N., 2005. Personalization for the Semantic Web. pp. 173–212, Springer-Verlag Berlin Heidelberg, 2005.
- Baldzer, J., Boll, S., Klante P., Krösche, J., Meyer, J., Rump, N., Scherp, A., Appelrath, H., 2004. Location-Aware Mobile Multimedia Applications on the Niccimon Platform. In 2. Braunschweiger Symposium – Informationssysteme für mobile Anwendungen (2004)

- Bankston, K., 2009, Facebook's New Privacy Changes: The Good, The Bad and The Ugly. [online] Available at:<<https://www.eff.org/deeplinks/2009/12/facebooks-new-privacy-changes-good-bad-and-ugly> > [Accessed 12 January 2010]
- Baron, N. S., 2007. My best day: Presentation of self and social manipulation in facebook and instant messaging. In Eighth International Conference, Association of Internet Researchers, 2007.
- BBC Online, 2007. Facebook Opens Profiles to Public. BBC Online.[online] Available at: <<http://news.bbc.co.uk/2/hi/technology/6980454.stm>> [Accessed 7 November 2007]
- Benelli, G., Bianchi, A., Marti, P. and Sennati, E., 1999. HIPS: Hyper-Interaction within Physical Space, IEEE, 1999.
- Bergman, M., 2011. Sweet Tools (PHP/ RDF).[online] Available at: <<http://www.mkbergman.com/sweet-tools/>> [Accessed 3 February 2011].
- Berners-Lee, T. and Shadbolt, N., 2009. Put in your postcode, out comes the data. The Times of London.
- Berners-Lee, T., 2006. Linked Data - Design Issues.[online] Available at:<<http://www.w3.org/DesignIssues/LinkedData.html>> [Accessed 10 January 2011]
- Berners-Lee, T., Chen, T. Y., Chilton, L., Connolly, D., Dhanaraj, R., Hollenbach, J., Lerer, A. and Sheets, D., 2006. Tabulator: Exploring and Analyzing Linked Data on the Semantic Web. Proceedings of the 3rd International Semantic Web User Interaction Workshop (SWUI06).
- Berners-Lee, T., 1998. Semantic Web Road Map. [Online]. Available: <http://www.w3.org/DesignIssues/Semantic.html>
- Berners-Lee, T., 2006. Isn't It Semantic? [Online]. Available: <http://www.ecs.soton.ac.uk/about/berners-lee.php>

- Bezerra B. L. D. and de Carvalho F. de A. T., 2004. A Symbolic Approach for Content-Based Information Filtering. *Information Processing Letters*, vol. 92(1), pp 45-52, October, 2004.
- Biddulph, M., 2005. Using Wikipedia and the Yahoo API to give structure to flat lists <http://www.hackdiary.com/2005/09/02/using-wikipedia-and-the-yahoo-api-to-give-structure-to-flat-lists/>, 2005.
- Billsus, D. and Pazzani, M., 2000. User Modeling for Adaptive News Access. *User-Modeling and User-Adapted Interaction* 10(2-3), 147-180.
- Bitirim, Y., Tonta, Y. and Sever, H., 2002, Information Retrieval Effectiveness of Turkish Search Engines, *Advances in Information Systems, Lecture Notes in Computer Science*, T. Yakhno (Ed.), vol. 2457, pp. 93-103, Springer-Verlag, Heidelberg, October 2002.
- Bizer, C. and Cyganiak, R., 2006. D2R Server - Publishing Relational Databases on the Semantic Web. Poster at the 5th International Semantic Web Conference (ISWC2006).
- Bizer, C., 2009. The Emerging Web of Linked Data. *Journal of Intelligent Systems*, IEEE. 2009
- Bizer, C., Heath, T., Berners-Lee, T., 2009. Linked Data— The Story So Far, *International Journal on Semantic Web and Information Special Issue on Linked Data*, 2009.
- Bobillo, F., Delgado, M., Gomez-Romero, J., 2008. Representation of context-dependant knowledge in ontologies: A model and an application. *Expert Systems with Applications*, 35, 1899-1908 (2008)
- Bojars, U., Breslin, J.G., Finn, A. and Decker, S., 2008. Using the semantic Web for linking and reusing data across Web2.0 communities. *Web Semantics: Science, Services and Agents on the World Wide Web* 6(1), 21–28 (2008)
- Bollmann P., Jochum R., Reiner U., Weissmann V. and Zuse H., 1986. Planung und Durchführung der Retrievaltests, *Leistungsbewertung von Information Retrieval*

- Verfahren, H. Scheider (Ed.), pp. 183-212, Fachbereich Informatik, Technische Universität Berlin, Germany, 1986.
- Börzsönyi, S., Kossmann, D., and Stocker, K., 2001. The Skyline Operator, Proceedings of the 17th International Conference on Data Engineering, p.421-430, April 02-06, 2001
- Boyd, D. and Ellison, N. B., 2007. Social network sites: Definition, history and scholar-ship. *Computer-Mediated Communication*, 1(13):11, 2007.
- Breese, J. S., Heckerman, D. and Kadie, C., 1998. Empirical Analysis of Predictive Algorithms for Collaborative Filtering, in Proceedings of the Fourteenth Conference on Uncertainty in Artificial Intelligence, pp. 43-52, 1998.
- Brickley, D. and Miller, L., 2005. FOAF Vocabulary Specification, working draft [online] Available at: < <http://xmlns.com/foaf/0.1> > [Accessed 10 April 2009]
- Bridge D. and Ferguson A., 2002. Diverse Product Recommendations using an Expressive Language for Case Retrieval. In Proceedings of ECCBR 2002: 6th European Conference on Advances in Case-Based Reasoning (LNCS 2416), pp. 43-57, 2002.
- Brusilovsky, P., 1996. Methods and techniques of adaptive hypermedia. *User Modeling and User Adapted Interaction*, 6(2-3):87–129, 1996.
- Brusilovsky, P., 2001. Adaptive hypermedia. *User Modeling and User-Adapted Interaction*, 11:87–110, 2001.
- Burke, R., 2002. Hybrid recommender systems: Survey and experiments. *User Modeling and User-Adapted Interaction*, 12:331–370, 2002.
- Buscaldi, D., Rosso, P., Arnal, E.S., 2005. A wordnet-based query expansion method for geographical information retrieval. In: Working Notes for the CLEF Workshop. (2005)
- Butuc, M.G., 2009. Modern PHP RDF toolkits:a comparative study.[online] Available at : <<http://www.slideshare.net/mariusbutuc/modern-php-rdf-toolkits-a-comparative-study>> [Accessed 10 December 2009]

- Calore, M., 2008. As Facebook Connect Expands, OpenID's Challenges Grow. *Wired*. December 1, 2008.
- Cantador, I., Szomszor, M., Alani, H., Fernández, M. and Castells, P., 2008. Enriching Ontological User Profiles with Tagging History for Multi-Domain Recommendations. In *CISWeb 1st International Workshop on Collective Semantics: Collective Intelligence & the Semantic Web*, (Tenerife, Spain, June, 2008).
- Cantador, I., Szomszor, M., Alani, H., Fernández, M. and Castells, P., 2008. Enriching Ontological User Profiles with Tagging History for Multi-Domain Recommendations. In *CISWeb 1st International Workshop on Collective Semantics: Collective Intelligence & the Semantic Web*, (Tenerife, Spain, June, 2008).
- Cao, H., Qi, Y., Candan, K. S., and Sapino, M. L. 2009. Exploring path query results through relevance feedback. In *Proceedings of the 18th ACM conference on Information and knowledge management (CIKM '09)*. ACM, New York, NY, USA, 1959-1962.
- Catarci, T., Dongilli, P., Mascio, T. D., Franconi, E., Santucci, G. and Tessaris, S., 2004. An ontology based visual tool for query formulation support. *Proceedings of the 16th European Conference on Artificial Intelligence*. IOS Press, Aug 2004, pages 308–312.
- Chan, C., Felber, P., Garofalakis, M., and Rastogi, R., 2002. Efficient filtering of xml documents with xpath expressions. In *Proceedings of the 18th International Conference on Data Engineering*, page 235244, 2002.
- Chen, D., and Wong, R. K., 2004. Optimizing the lazy DFA approach for XML stream processing. In *Proceedings of the 15th Australasian database conference - Volume 27 (ADC '04)*, Australia, 131-140.
- Charnigo, L. and Barnett-Ellis, P., 2007. Checking out facebook.com: The impact of a digital trend on academic libraries. *Information Technology and Libraries*, 26(1): 23, 2007.

- Cheng, G. and Qu, Y., 2009. Searching Linked Objects with Falcons: Approach, Implementation and Evaluation. *International Journal on Semantic Web and Information Systems*, Special Issue on Linked Data. 2009.
- Cheverst, K., Davies, N., Mitchell, K., Friday, A., Efstratiou, C., 2002. Developing a contextaware electronic tourist guide: some issues and experiences. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (The Hague, The Netherlands, April 01 - 06, 2000)*. CHI '00. ACM Press. (2002)
- Cheverst, K., Davies, N., Mitchell, K., Friday, A.: Experiences of developing and deploying a context-aware tourist guide: the GUIDE project, *Proceedings of the 6th annual international conference on Mobile computing and networking*, August, Boston. (2000)
- Chun, S., Cherry, R., Hiwiller, D., Trant, J. and Wyman, B., 2006. Steve.museum: An Ongoing Experiment in Social Tagging, Folksonomy and Museums. *Museums and the Web Conference*. Albuquerque, March, 2006.
- Ciavarella, C. and Paterno, F., 2004. The design of a handheld, location-aware guide for indoor environments. *Personal Ubiquitous Computing* (2004) 8, pages 82–91, 2004.
- Claypool, M., Gokhale, A., Miranda, T., Murnikov, P., Netes, D. and Sartin, M., 1999. Combining Content-Based and Collaborative Filters in an Online Newspaper. *SIGIR '99 Workshop on Recommender Systems: Algorithms and Evaluation*. Berkeley, CA.
- Coetzee, P., Heath, T. and Motta, E., 2008. SparqPlug. *Proceedings of the 1st Workshop on Linked Data on the Web (LDOW2008)*.
- Cooper, W. B., 1998, Expected Search Length: A Single Measure of Retrieval Effectiveness Based on the Weak Ordering Action of Retrieval Systems, *American Documentation*, vol. 19, pp. 30-41, 1998.

Cubrilovic, N., 2007. Facebook Source Code Leaked. [online]TechCrunch, August 11, 2007. Available at:<<http://techcrunch.com/2007/08/11/facebook-source-code-leaked/>> [Accessed 10 August 2009]

Cyganiak, R., Bizer, C., 2008. Pubby - A Linked Data Frontend for SPARQL Endpoints. [online] Available at: < <http://www4.wiwiiss.fu-berlin.de/pubby/>> [Accessed 14 June 2009]

Davis, I., 2009. If you love something... set it free. [online] Available at:<<http://slidesha.re/haCax>>, 2009, keynote speech at Code4Lib. [Accessed 10 March 2010]

Domingos, P., 2005. Mining social networks for viral marketing. IEEE Intelligent Systems, 20(1):80–82, 2005.

Dumbill, E., 2002a. Finding friends with xml and rdf, IBM's XML Watch, [online] Available at: <<http://www-106.ibm.com/developerworks/xml/library/x-FOAF.html>>, [Accessed 10 July 2009].

Dumbill, E., 2002b. "Support online communities with FOAF: How the friend-of-a-friend vocabulary addresses issues of accountability and privacy," IBM's XML Watch, <http://www-106.ibm.com/developerworks/xml/library/x-FOAF2.html>, August 2002.

Dumbill, E., 2003. Tracking provenance of rdf data, IBM's XML Watch, [online] Available at: <<http://www-106.ibm.com/developerworks/xml/library/x-rdfprov.html>>, [Accessed 10 July 2009].

Dwyer, C., Hiltz, S. R. and Passerini, K., 2007. Digital relationships in the 'myspace' generation: Results from a qualitative study. In Proceedings of AMCIS 2007, 2007. Ermecke, R., Mayrhofer, P. and Wagner, S., 2009. Agents of diffusion insights from a survey of facebook users. In Proceedings of the Forty-second Hawaii International Conference on System Sciences (HICSS-2007), Los Alamitos, CA, 2009.

Eulerfx, 2009. [online] Available at: <http://stackoverflow.com/questions/307291/how-does-the-google-did-you-mean-algorithm-work>, [Accessed 2012]

- Facebook , 2010a. Facebook. [online] Available at:  
 <<http://en.wikipedia.org/wiki/Facebook>>, [Accessed 29 September 2010]
- Facebook, 2010b. Facebook Factsheet. [online] Available at:  
 <<http://www.facebook.com/press/info.php?factsheet>>, [Accessed 29 September 2010]
- Facebook Graph, 2012. Graph API. [online] Available at:  
<http://developers.facebook.com/docs/reference/api/>, [Accessed 2011, 2012]
- Fan, H. and Poole, M. S., 2006. What is personalization? Perspectives on the design and implementation of personalization in information systems. *Journal of Organizational Computing and Electronic Commerce*. 16 (3 & 4), pp. 179 – 202, 2006.
- Felt, A. 2007. Defacing Facebook: A Web 2.0 Case Study. [online] Available at:<[www.cs.virginia.edu/felt/fbook/felt-facebook.pdf](http://www.cs.virginia.edu/felt/fbook/felt-facebook.pdf)> [Accessed 20 June 2008]
- Felt, A., Hooimeijer, P., Evans, D. and Weimer, W., 2008. Talking to strangers without taking their candy: isolating proxied content. In *Proceedings of the 1st Workshop on Social Network Systems (SocialNets '08)*. ACM, New York, USA, 25-30. DOI=10.1145/1435497.1435502  
<http://doi.acm.org/10.1145/1435497.1435502>
- Fikes, R., Hayes, P. and Horrocks, I., 2003. OWL-QL: A language for deductive query answering on the semantic Web. Technical Report, Knowledge Systems Laboratory, Stanford University, Stanford, CA, 2003.
- Fleck, M., Frid, M., Kindberg, T., Rajani, R., O'Brien- Strain, E. and Spasojevic, M., 2002. From Informing to Remembering: Deploying a Ubiquitous System in an Interactive Science Museum. *Pervasive Computing* 1(2): pages 13- 21, 2002.
- Frank, E., Paynter, G. W., Witten, I. H., Gutwin, C. and Nevill-Manning, C. G., 1999. Domain-specific keyphrase extraction. In *IJCAI*, 1999.



- Frivolt, G. and Bielikova, M., 2006. Growing World Wide Social Network by Bridging Social Portals Using FOAF. In 15th Int. Conf. on Knowledge Engineering and Knowledge Management Managing Knowledge in a World of Networks, EKAW 2006, Poster, pages 9 10, 2006.
- Frivolt, G., Bieliková, M., 2007. Ensuring privacy in FOAF profiles. in Znalosti (2007)
- Ganu, G., Elhadad, N. and Marian, A., 2009. Beyond the Stars: Improving Rating Predictions using Review Text Content. Twelfth International Workshop on the Web and Databases (WebDB 2009), June 28, 2009, Providence, Rhode Island, USA. (Ganu, et al., 2009)
- Ganu, G., Kakodkar, Y. and Marian, A., 2012. Improving the quality of predictions using textual information in online user reviews. Information Systems (March 2012) (Ganu, et al., 2012) [12 13]
- Giles, J., 2005. Internet encyclopaedias go head to head. Nature, 438(7070):900-901, December 2005.
- Golbeck, J. and Hendler, J., 2006. Inferring binary trust relationships in Web- based social networks. ACM Trans. Internet Technol., 6(4):497–529, 2006.
- Golder, S.A. and Huberman, B.A., 2006. Usage patterns of collaborative tagging systems. Journal of Information Science 32, 198–208 (2006)
- Grinter R. E., Aoki, P. M., Szymanski M. H., Thornton J. D., Woodruff, A., 2002. Revisiting the Visit: Understanding How Technology can shape the Museum Visit, ACM, 2002.
- Gross, R. and Acquisti, A., 2005. Information revelation and privacy in online social networks. In Workshop on Privacy in the Electronic Society, Alexandria, VA, 2005.
- Gruber, T., 2008. Collective Knowledge Systems: Where the Social Web meets the Semantic Web. Journal of Web Semantics 6(1) (2008)

- Guha, R., McCool, R. and Miller, E., 2003. Semantic search. WWW'03: Proceedings of the 12th international conference on World Wide Web. ACM Press, 2003, pages 700–709.
- Hall, W., 2011. The Ever Evolving Web: The Power of Networks. *International Journal of Communication* 5 (2011), 651-664 (2011).
- Hampton, K., Goulet, L. S., Rainie, L. and Purcell, K., 2011. Social networking sites and our lives. Technical Report, Pew Internet & American Life Project, 2011.
- Haslhofer, B., Schandl, B., 2008. The OAI2LOD Server: Exposing OAI-PMH Metadata as Linked Data. Proceedings of the 1st Workshop about Linked Data on the Web (LDOW2008).
- Hassanzadeh, O., Lim, L., Kementsietsidis, A., Wang, M., 2009. A Declarative Framework for Semantic Link Discovery over Relational Data. Poster at 18th World Wide Web Conference (WWW2009).
- Heflin, J., Hendler, J., 2000. Searching the Web with SHOE. Artificial Intelligence for Web Search, Papers from the workshop, AAAI 2000. AAAI Press, 2000, pages 35–40
- Hendler, J., Shadbolt, N., Hall, W., Berners-Lee, T., and Weitzner, D., 2008. Web science: an interdisciplinary approach to understanding the Web. *Commun. ACM* 51, 7 (July 2008).
- Henze, N. and Nejd, W., 2000. Extendible adaptive hypermedia courseware: Integrating different courses and Web material. In Proceedings of the International Conference on Adaptive Hypermedia and Adaptive Web-Based Systems (AH 2000), Trento, Italy, 2000.
- Herlocker, J. L., Konstan, J. A., Terveen, L. G. and Riedl, J. T., 2004. Evaluating Collaborative Filtering Recommender Systems, *ACM Trans. Inf. Syst.*, vol. 22(1), pp. 5-53, 2004.
- Hill, W., Stead, L., Rosenstein, M. and Furnas, G., 1995. Recommending and evaluating choices in a virtual community of use. In CHI '95: Conference

Proceedings on Human Factors in Computing Systems, Denver, CO, pp. 194-201.

Hiltz, R. and Turoff, M., 2003. *The Network Nation 2. S.* - Addison-Wesley Educational Publishers Inc. March 1979-2003.

Hjorland, B. and Christensen, F. S., 2002. Work tasks and socio-cognitive relevance: A specific example. *J. Am. Soc. Inf. Sci. Technol.* 53, 11 (Sep. 2002), 960-965. DOI= <http://dx.doi.org/10.1002/asi.10132>.

Hogan, A., Harth, A., Umrich, J. and Decker, S., 2007. Towards a scalable search and query engine for the Web. *Proceedings of the 16th Conference on World Wide Web (WWW2007)*.

Hristova, N., O'Hare, G. M. P., Lowen, T., 2003. Agent-based ubiquitous systems: 9 lessons learnt. In *Workshop on System Support for Ubiquitous Computing (UbiSys'03)*, 5th International Conference on Ubiquitous Computing (UbiComp), Seattle, WA, USA (2003)

Hsu, E. I. and McGuinness, D. L., 2003. Wine Agent: Semantic Web testbed application. *Description Logics*, Calvanese, D., Giacomo, G. D. and Franconi, E., editors, volume 81 of *CEUR Workshop Proceedings*. CEUR-WS.org, 2003.

Huang Z., Chen H. and Zeng D., 2004. Applying Associative Retrieval Techniques to Alleviate the Sparsity Problem in Collaborative Filtering. *ACM Trans. Inf. Syst.*, vol. 22(1), pp. 116-142, 2004

Huynh, T. D., Jennings, N. R. and Shadbolt, N. R., 2006. An integrated trust and reputation model for open multi-agent systems. *Journal of Autonomous Agents and Multi-Agent Systems*, 13(2):119–154, 2006.

Hyvönen, E., Mäkelä, E., Salminen, M., Valo, A., Viljanen, K., Saarela, S., Junnila, M. and Kettula, S., 2005. MuseumFinland - Finnish Museums on the Semantic Web. *Journal of Web Semantics*, vol. 3, no. 2, pp. 25, 2005

- IEEE. IEEE P1484.2/D7, 2000-11-28. draft standard for learning technology. Public and private information (papi) for learners (papi learner). [online] <<http://ltsc.ieee.org/wg2/>> [Retrieved 19<sup>th</sup> December, 2011]
- IMS LIP, 2005. IMS Learner Information Package Specification (LIP). 2005 [online] <[www.imsglobal.org](http://www.imsglobal.org)> [Retrieved 19<sup>th</sup> December, 2011].
- Iturrioz, J., Diaz, O. and Arellano, C., 2007. Towards federated Web2.0 sites: The tagmas approach. In Tagging and Metadata for Social Information Organization Workshop, WWW07, 2007.
- Järvelin, K. and Kekäläinen J., 2002. Cumulated Gain-Based Evaluation of IR Techniques, ACM Transaction Information Systems, Oct. 2002, pp. 422-446.
- Joachims, T., 1997. A Probabilistic Analysis of the Rocchio Algorithm with TFIDF for Text Categorization. Proceedings of the Fourteenth International Conference on Machine Learning, p.143-151, July 08-12, 1997.
- Jones, H., Soltren, J. H., 2005. Facebook: Threats to Privacy. Cambridge, MA: MIT (MIT 6.805/STS085: Ethics and Law on the Electronic Frontier – 14th December 2005).
- Kamar, A., 2003. Mobile Tourist Guide (m-ToGuide). Deliverable 1.4, Project Final Report. IST-2001-36004 (2003)
- Karger, D. R., Bakshi, K., Huynh, D., Quan, D. and Sinha, V., 2005. Haystack: A general purpose information management tool for end users based on semistructured data. Proceedings of the CIDR Conference, 2005, pages 13– 26.
- Kauppinen, T., Henriksson, R., Sinkkilä A. R., Lindroos, R., Vaatainen, J., Hyvonen, E., 2008. Ontology-based disambiguation of spatiotemporal locations. In: 1st international workshop on Identity and Reference on the Semantic Web (IRSW2008), 5th European Semantic Web Conference 2008 (ESWC 2008), Tenerife, Spain. (June1-5 2008).

- Khalefa, M. E., Mokbel, M. F., Levandoski, J. J., 2008. Skyline Query Processing for Incomplete Data, Proceedings of the 2008 IEEE 24th International Conference on Data Engineering, p.556-565, April 07-12, 2008
- Kim H., Chan P., 2003. Learning Implicit User Interest Hierarchy for Context in Personalization. In Proceedings of the 2003 International Conference on Intelligent user interfaces 2003, Miami, Florida.
- Kim, H., Yang, S., Song, S., Breslin, J. G. and Kim, H., 2007. Tag Mediated Society with SCOT Ontology. ISWC2007. (2007).
- Kim, J. W., and Candan, S. K., 2009. Skip-and-prune: cosine-based top-k query processing for efficient context-sensitive document retrieval. In Proceedings of the 35th SIGMOD international conference on Management of data (SIGMOD '09), ACM, New York, NY, USA, 115-126.
- Kochut, K., Janik, M. 2007. SPARQLer: Extended SPARQL for semantic association discovery. ESWC 2007, Lecture Notes in Comput. Sci., vol. 4519 (2007), pp. 145–159
- Koshizuka, N. and Sakamura, K., 2000. The Tokyo University Museum. Kyoto International Conference on Digital Libraries: Research and Practice, November 13 - 16, 2000.
- Kossinets, G. and Watts, D. J., 2009. Origins of homophily in an evolving social network. Am. J. Sociol., 115(2):405-450, September 2009.
- Krulwich, B., 1997. Lifestyle Finder: Intelligent User Profiling Using Large-Scale Demographic Data. Artificial Intelligence Magazine 18 (2), 37-45.
- Kruse, P. M., Naujoks, A., Roesner, D. and Kunze, M., 2005. Clever Search: A WordNet based wrapper for internet search engines. Proceedings of the 2nd GermaNet Workshop, 2005.
- Lam W. and Mostafa J., 2001. Modeling user interest shift using a bayesian approach. J. Am. Soc. Inf. Sci., 52: 416–429.(Lam and Mostafa, 2001)

- Lam X. N., Vu T., Le T. D., and Duong A. D. 2008. Addressing cold-start problem in recommendation systems. In Proceedings of the 2nd international conference on Ubiquitous information management and communication (ICUIMC '08). ACM, New York, NY, USA.
- Lang, K., 1995. Newsweeder: Learning to Filter news. In Proceedings of the 12th International Conference on Machine Learning, Lake Tahoe, CA, pp. 331-339.
- Li, W. S., Candan, K. S., Hirata, K., Hara, Y., 2001. Supporting efficient multimedia database exploration. VLDB J. 9(4): 312-326 (2001)
- Li, X., Guo, L. and Zhao, Y.E., 2008. Tag-based social interest discovery. In: Proc. 19th Int. World Wide Web Conf. (WWW), (Beijing, China 2008)
- Li, X., Guo, L., Zhao, Y.E. 2008. Tag-based social interest discovery. In: Proc. 19th Int. World Wide Web Conf. (WWW), (Beijing, China 2008)
- Li, Z., Zhou, D., Juan, Y.F., and Han, J. 2010. Keyword extraction for social snippets. In Proceedings of the 19th international conference on World wide Web (WWW '10). ACM, New York, NY, USA, 1143-1144.
- Luyten, K. and Coninx, K., 2004. ImogI: Take Control over a Context Aware Electronic Mobile Guide for Museums. HCI in Mobile Guides, 13 September 2004, University of Strathclyde, Glasgow.
- Maedche, A., Staab, S., Stojanovic, N., Studer, R. and Sure, Y., 2001. SEAL — a framework for developing semantic Web portals. Advances in Databases, Proceedings of the 18th British National Conference on Databases, Jul 2001, pages 1–22.
- Marlow, C., Naaman, M., Boyd, D., Davis, M., 2006. HT06, tagging paper, taxonomy, flickr, academic article, to read. In Proceedings of International Conference on Hypertext (Odense, Denmark 2006)
- Mase, K., Sumi, Y. and Kadobayashi, R., 2000. The Weaved Reality: What Context-aware Interface Agents Bring About. Invited Session at Asian Conference on Computer Vision. ACCV2000, Jan 2002, Taipei.

- Mathes, A. Folksonomies - Cooperative Classification and Communication Through Shared Metadata. UIC Technical Report, 2004.
- McCarthy, C., 2009. Facebook backtracks on public friend lists. CNN News, December 11, 2009. [http://news.cnet.com/8301-13577\\_3-10413835-36.html](http://news.cnet.com/8301-13577_3-10413835-36.html)
- McNee S. M., Lam S. K., Guetzlaff C., Konstan J. A. and Riedl J., 2003a. Confidence Displays and Training in Recommender Systems. In Proceedings of the INTERACT '03 IFIP TC13 International Conference on Human-Computer Interaction, pp. 176-183, 2003. (McNee, et al., 2003a)
- McNee S. M., Lam S. K., Konstan J. A. and Riedl J., 2003b. Interfaces for Eliciting New User Preferences in Recommender Systems, in Proceedings of the 9th International Conference on User Modeling (UM'2003), pp. 178-187, 2003.
- McNee S. M., Riedl J. and Konstan J. A., 2006. Being accurate is not enough: how accuracy metrics have hurt recommender systems. In CHI '06 extended abstracts on Human factors in computing systems (CHI EA '06). ACM, New York, NY, USA, 1097-1101. (McNee, et al., 2006)
- McNee, S. M., 2006, Meeting User Information Needs in Recommender Systems. PhD thesis, University of Minnesota- Twin Cities, June 2006. [55]
- McNee, S. M., Riedl, J. and Konstan, J. A., 2006. Making Recommendations Better: An Analytic Model for Human-Recommender Interaction, in Extended Abstracts of the 2006 ACM Conference on Human Factors in Computing Systems (CHI 2006), pp. 1003-1008, 2006.
- McNee, S. M., Riedl, J. and Konstan, J. A., 2006. Being Accurate is Not enough: How Accuracy Metrics have Hurt Recommender Systems, in Extended Abstracts of the 2006 ACM Conference on Human Factors in Computing Systems (CHI 2006), pp. 997-1001, 2006.
- Melville. P., Mooney, R. J. and Nagarajan, R., 2002. Content-Boosted Collaborative Filtering for Improved Recommendations, Proceedings of the Eighteenth National Conference on Artificial Intelligence(AAAI-2002),pp. 187-192, Edmonton, Canada, July 2002

- Mika, P., 2004. Social Networks and the Semantic Web. In IEEE/WIC/ACM Int. Conf. on Web Intelligence, WI 2004, pages 285{291, Los Alamitos, CA, USA, 2004. IEEE Computer Society.
- Miles, Alistair and Dan Brickley, 2004. SKOS Mapping Vocabulary Specification. W3C.[online] Available at:  
<<http://www.w3.org/2004/02/skos/mapping/spec/2004-11-11.html>> [Accessed 10 April 2009].
- Miller, G. A., 1995, Wordnet: a lexical database for english. Communications. ACM, 38(11):39-41, 1995.
- Moldovan, D. I. and Mihalcea, R., 2000. Using WordNet and lexical operators to improve internet searches. IEEE Internet Computing, 4,1(2000), pages 34–43.
- Murnan, C. A., 2006. Expanding communication mechanisms: they're not just e-mailing anymore. In SIGUCCS '06: Proceedings of the 34th annual ACM SIGUCCS conference on User services, pages 267–272, New York, NY, USA, 2006.
- Nasirifard, P., Michael, H. M. and Decker, S., 2009, Privacy Concerns of FOAF-Based Linked Data. Trust and Privacy on the Social and Semantic Web, The 6th Annual European Semantic Web Conference (ESWC2009). Heraklion, Greece June 1st 2009.
- Noor, S. and Martinez, K., 2009, Using social data as context for making recommendations: an ontology based approach. In Proceedings of the 1st Workshop on Context, Information and Ontologies. Heraklion, Greece, June 01, 2009.
- Nowack, B., 2009, RDF and SPARQL for PHP Developers.[online] Available at:  
<<http://slidesha.re/eHQtn>> at New York Semantic Web Meetup. [Accessed on 13 March 2010]
- Nowack, B., 2009, Workshop - Easy RDF and SPARQL for LAMP systems – Building Semantic Web Apps on PHP & MySQL, Friday, May 22, 2009 from 9:00 AM to 1:00 PM (ET) New York, NY.



- O'Grady, M.J., O'Hare G.M.P., 2004. Gulliver's Genie: Agency, Mobility & Adaptivity. Computers & Graphics. Special Issue on Pervasive Computing and Ambient Intelligence - Mobility, Ubiquity and Wearables Get Together, Vol. 28, No. 4, Elsevier. (2004)
- Oppermann, R. and Specht, M., 2000. A Context-sensitive Nomadic Information System as an Exhibition Guide. Proceedings of the Handheld and Ubiquitous Computing Second International Symposium, HUC 2000, Bristol, UK, September 25-27, pages 127 – 142, 2000.
- Pauls, A. D. and Klein, 2011. Faster and Smaller N-Gram Language Model. in Proceeding HLT '11 Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1, 2011.
- Passant, A. and Laublet, P., 2008. Meaning Of A Tag: A collaborative approach to bridge the gap between tagging and Linked Data. In Proceedings of the WWW 2008 Workshop Linked Data on the Web (LDOW2008),( Beijing, China, Apr 2008).
- Patil, S. and Lai, J., 2005, Who gets to know what when: configuring privacy preferences in an awareness application. In ACM Conference on Human Factors and Computing Systems, Portland, OR, 2005.
- Paul J., Getty Trust, 2010. Art & Architecture Thesaurus (AAT). [online] Available at:  
<[http://www.getty.edu/research/conducting\\_research/vocabularies/aat/about.html](http://www.getty.edu/research/conducting_research/vocabularies/aat/about.html)  
l> [Accessed on 4 April 2010]
- Pazzani, M. J., 1999, A Framework for Collaborative, Content-Based and Demographic Filtering. Artificial Intelligence Review, 13 (5/6), 393-408.
- Pennock, D. M., Horvitz, E. and Giles, C. L., 2000, Social choice theory and recommender systems: Analysis of the axiomatic foundations of collaborative Filtering. In AAAI/IAAI, pages 729-734, 2000.

- PHP XML Classes, 2002, A collection of classes and resources to process XML using PHP [online] Available at:<<http://phpxmlclasses.sourceforge.net/>> [Accessed 24 March 2009][4]
- Poslad, S., Laamanen, H., Malaka, R., Nick, A., Buckle, P. and Zipf, A., 2001. CRUMPET: Creation of User-Friendly Mobile Services Personalised for Tourism. In: Proceedings of: 3G 2001 - Second Int. Conf. on 3G Mobile Communication Technologies.2001.
- Qi, Y., Candan, K. S., and Sapino, M. L., 2007. Sum-max monotonic ranked joins for evaluating top-k twig queries on weighted data graphs. In Proceedings of the 33rd international conference on Very large data bases (VLDB '07). VLDB Endowment 507-518.
- Quan, D., Huynh, D. and Karger, D. R., 2003, Haystack: A platform for authoring end user semantic Web applications. Proceedings of the Second International Semantic Web Conference, 2003, pages 738–753.
- Rao, R., Pedersen, J. O., Hearst, M. A., Mackinlay, J. D., Card, S. K., Masinter, L., Halvorsen P. and Robertson, G. C., 1995, Rich Interaction in the Digital Library, Comm. ACM, Apr. 1995, pp. 29-39.
- RAP, 2008. RDF API for PHP V0.9.6[online] Available at:<<http://www4.wiwiwiss.fu-berlin.de/bizer/rdfapi/index.html> > [Accessed 15 May 2011].
- Rashid A. M. ,Albert I., Cosley D., Lam S. K., McNee S. M., Konstan J. A. and Riedl J., 2002. Getting to Know You: Learning New User Preferences in Recommender Systems. in Proceedings of the 7th International Conference on Intelligent User Interfaces, pp. 127-134, 2002.
- Rahman M.M., Davis D.N., 2012. "Fuzzy Unordered Rules Induction Algorithm Used as Missing Value Imputation Methods for K-Mean Clustering on Real Cardiovascular Data". Proceedings of The World Congress on Engineering 2012 **1** (1): 391–394

- Rashid, A. M., Albert, I., Cosley, D., Lam, S. K., McNee, S. M., Konstan, J. A. and Riedl, J., 2002, Getting to Know You: Learning New User Preferences in Recommender Systems, in Proceedings of the 7th International Conference on Intelligent User Interfaces, pp. 127-134, 2002.
- Resnick, P. and Varian, H. R., 1997, Recommender Systems. Communications of the ACM, 40 (3), 56-58.
- Resnick, P., Iacovou, N., Suchak, M., Bergstrom, P. and Riedl, J., 1994, GroupLens: An Open Architecture for Collaborative Filtering of Netnews. In Proceedings of the Conference on Computer Supported Cooperative Work, Chapel Hill, NC, pp. 175-186.
- Ritzer, G. & Jurgenson, N., 2010, Production, Consumption, Prosumption. Journal of Consumer Culture, 10(1), pp. 13 –36. 2010.
- Rocchi, C., Stock, O., Zancanaro, M., Kruppa, M. And Krüger, A. 2004, The Museum Visit: Generating Seamless Personalised Presentations on Multiple Devices. In Proceedings of the Intelligent User Interfaces 2004, January 13-16, 2004 Island of Madeira, Portugal.(PEaCH)
- Rocchio, J., 1971. Relevance feedback in information retrieval. In The SMART Retrieval System (1971), pp. 313-323. Rocchi, C., Stock, O., Zancanaro, M., Kruppa, M. And Krüger, A., 2007, Adaptive Intelligent Presentation of Information for the Visitors, in PEACH, Springer, 2007.
- Rocha, C., Schwabe, D. and de Aragão, M. P., 2004, A hybrid approach for searching in the semantic Web. Proceedings of the 13th international conference on World Wide Web, May 2004, pages 374–383.
- Rosenblum, D., 2007, What anyone can know: The privacy risks of social networking sites. Security and Privacy, IEEE, 5(3):40–49, 2007.
- Roussopoulos, N., Kelley, S., and Vincent, F., 1995. Nearest neighbour queries. SIGMOD Rec. 24, 2 (May 1995), 71-79.

- Ruthven, I., and Lalmas, M. 2003. A survey on the use of relevance feedback for information access systems. *Knowl. Eng. Rev.* 18, 2 (June 2003), 95-145.
- Rutledge, L., Aroyo, L. and Stash, N., 2006, Determining user interests about museum collections. In *Proc. the 15th International Conference on World Wide Web (WWW'06)* Edinburgh, Scotland, 2006.
- Sarwar, B. M., Konstan, J. A., Borchers, A., Herlocker, J. Miller, B. and Riedl, J., 1998, Using Filtering Agents to Improve Prediction Quality in the GroupLens Research Collaborative Filtering System. In *Proceedings of the ACM 1998 Conference on Computer Supported Cooperative Work*, Seattle, WA, pp. 345-354.
- Schafer, J. B., Konstan, J. and Riedl, J., 1999, Recommender Systems in E-Commerce. In *EC '99: Proceedings of the First ACM Conference on Electronic Commerce*, Denver, CO, pp. 158-166.
- Schein, A. I., Popescul, A., Ungar, L. H. and Pennock, D. M., 2002, Methods and metrics for cold-start recommendations. In *25th International ACM Conference on Research and Development in Information Retrieval*, 2002.
- Scherp, A., Boll, S., 2004. mobileMM4U-framework support for dynamic personalised multimedia content on mobile systems. In *Proc. des Techniques and Applications for Mobile Commerce (TaMoCO) Track der Multi-Konferenz Wirtschaftsinformatik 2004*, Essen, Deutschland, März 2004, 3, Aka GmbH, S.204-215 (2004)
- Schmidt-Belz, B., Poslad, S., Nick, A., Zipf, A., 2002. Personalized and Location-based Mobile Tourism Services. *Workshop on Mobile Tourism Support Systems in conjunction with Mobile HCI '02*. Pisa. (2002)
- Schwab, I., Kobsa, A. and Koychev, I., 2001, Learning User Interests through Positive Examples Using Content Analysis and Collaborative Filtering. Internal Memo, GMD, St. Augustin, Germany.

- Schwab, I., Kobsa, A. and Koychev, I., 2001, Learning User Interests through Positive Examples Using Content Analysis and Collaborative Filtering. Internal Memo, GMD, St. Augustin, Germany. 2001.
- Scoble, R., 2008, Facebook disabled my account. Scobleizer. January 3, 2008.
- Semitsu J. P., 2011. From Facebook to Mug Shot: How the Dearth of Social Networking Privacy Rights Revolutionized Online Government Surveillance, Pace Law Review, Volume 31, Issue 1 Social Networking and the Law, Winter 2011.
- SFMOMA, 2001. San Francisco Museum of Modern Art. Points of Departure. [online] Available at: <<http://www.sfmoma.org/press/pressroom.asp?arch=y&id=117&do=events>> [Accessed 10 March 2007]
- Shadbolt, N., Hall, W. and Berners-Lee, T., 2006, The Semantic Web revisited. IEEE Intelligent Systems 2006, 21(3), 96–101
- Shannon, V., 2006. A ‘More Revolutionary’ Web. The New York Times [online] <<http://www.nytimes.com/2006/05/23/technology/23iht-Web.html>> [Retrieved, 12<sup>th</sup> February, 2011]
- Shardanand, U. and Maes, P., 1995, Social Information Filtering: Algorithms for Automating ‘Word of Mouth’, CHI ’95: Conference Proceedings on Human Factors in Computing Systems, Denver, CO, pp. 210-217.
- Sherry, H., 2002. The Electronic Guidebook: A study of User Experiences using Mobile Web Content in a Museum Setting. International Workshop on Whireless and Mobile Technologies in Education, IEEE, 2002.
- Sinclair, P., Lewis, P. and Martinez, K., 2007. Dynamic Link Service 2.0: using Wikipedia as a linkbase. In: Hypertext 2007, (Manchester, United Kingdom, 10 - 12 September 2007
- Smyth, B. and Cotter, P., 2000, A Personalized TV Listings Service for the Digital TV Age. Knowledge-Based Systems 13: 53-59.

- Snyder, J., Carpenter, D. and Slauson, G. J., 2006, Myspace.com - a social networking site and social contract theory. Proceedings of ISECON, 2006.
- Song, F. and Croft, W. B., (1999), A General Language Model for Information Retrieval, In CIKM '99: Proceedings of the eighth international conference on Information and knowledge management (1999), pp. 316-321
- Specht, M., 1998, Empirical evaluation of adaptive annotation in hypermedia. In EDMedia and ED-Telekom, Freiburg, Germany, 1998.
- Specia, L., Motta, E., 2007. Integrating folksonomies with the semantic Web. In ESWC 2007. LNCS, vol. 4519. Springer, Heidelberg (2007)
- Staab, S., Domingos, P., Mika, P., Golbeck, J., Ding, L., Finin, T., Joshi, A., Nowak, A., and Vallacher, R. R., 2005. Social Networks Applied. IEEE Intelligent Systems, 20(1):80-93, 2005.
- Strijbos, J., Martens, R., Prins, F., Jochems, W., 2006. Content analysis: What are they talking about? Computers & Education 46: 29–48.  
doi:10.1016/j.compedu.2005.04.002
- Stutzman, F., 2006, An evaluation of identity-sharing behavior in social network communities. Journal of the International Digital Media and Arts Association, (3): 10–18, 2006.
- Su, X., Khoshgoftaar, T. M. and Greiner (2008) Imputed Neighborhood Based Collaborative Filtering, In Web Intelligence (2008), pp. 633-639 (Su, et al., 2008)
- Szomszor, M., Alani, H., Cantador, I., O'Hara, K. and Shadbolt, N. 2008. Semantic Modelling of User Interests based on Cross-Folksonomy Analysis. In ISWC 7th International Semantic Web Conference (Karlsruhe, Germany, October 26 - 30, 2008).
- Szomszor, M., Alani, H., Cantador, I., O'Hara, K. and Shadbolt, N., 2008. Semantic Modelling of User Interests based on Cross-Folksonomy Analysis. In ISWC 7th

International Semantic Web Conference (Karlsruhe, Germany, October 26 - 30, 2008).

Tao, Y., Hristidis, V., Papadias, D., and Papakonstantinou, Y., 2007. Branch-and-bound processing of ranked queries. *Inf. Syst.* 32, 3 (May 2007), 424-445.

Teevan, J., Alvarado C., Ackerman, M. S. and Karger, D. R., 2004, The perfect search engine is not enough: a study of orienteering behavior in directed search. *Proceedings of the Conference on Human Factors in Computing Systems, CHI*, April 2004, pages 415–422.

Terveen, L. and Hill, W., 2001, Human-Computer Collaboration in Recommender Systems. In J. Carroll (ed.): *Human Computer Interaction in the New Millenium*. New York: Addison-Wesley, 487-509.

The Telegraph, 2007. Top 100 living geniuses. [online] Available at <http://www.telegraph.co.uk/news/uknews/1567544/Top-100-living-geniuses.html> [ Accessed 6<sup>th</sup> November, 2011]

Torres, R., McNee, S. M., Abel, M., Konstan, J. A. and Riedl, J., 2004, Enhancing Digital Libraries with TechLens+, in *Proceedings of the 2004 Joint ACM/IEEE Conference on Digital Libraries*, pp. 228-236.

Trant, J., 2009, Studying Social Tagging and Folksonomies: A Review and Framework. *Journal of Digital Information*, Volume 10, No. 1.

Tsatsou, D., Menemenis, F., Kompatsiaris, I. and Davis, P. C., 2009, A semantic framework for personalized ad recommendation based on advanced textual analysis, *RecSys '09 Proceedings of the third ACM conference on Recommender systems*.

Tunkelang, D., 2009, The Noisy Channel: Precision and Recall, [online] Available at: <http://thenoisychannel.com/2009/03/17/precision-and-recall/> [Accessed 10th March 2010]

Turney, P. D., 2000. Learning algorithms for keyphrase extraction. *Information Retrieval*, 2000.

- Uszkoreit, H., Xu, F., Liu, W., Steffen, J., Aslan, I., Liu, J., Muller, C., Holtkamp, B. and Wojciechowski, M., 2007. A successful field test of a mobile and multilingual information service system COMPASS2008. In Proceedings of the 12th international conference on Human-computer interaction: applications and services (HCI'07), Julie A. Jacko (Ed.). Springer-Verlag, Berlin, Heidelberg, 1047-1056.
- Van, G. L., Tuytelaars, T. and Pollefeys, M., 1999. Adventurous tourism for couch potatoes. Invited, Proc. CAIP99, LNCS 1689, Springer-Verlag, pp.98-107, 1999.
- Vladimir, I. L., 1996. Binary codes capable of correcting deletions, insertions and reversals. Soviet Physics Doklady, Vol. 10, No. 8. (1966), pp. 707-710)
- Volz, J., Bizer, C., Gaedke, M., Kobilarov, G., 2009. Silk – A Link Discovery Framework for the Web of Data. Proceedings of the 2nd Workshop on Linked Data on the Web (LDOW2009).
- Walter, C., 2009, Facebook's New Terms Of Service: We Can Do Anything We Want With Your Content. Forever. Consumerist, Consumer Media LLC, February 15, 2009.
- Wang, J. and Zeng, Y., 2011, Efficient mining of weighted frequent pattern over data streams, Fuzzy Systems and Knowledge Discovery (FSKD), 2011 Eighth International Conference on Issue Date: 26-28 July 2011, Volume: 2, pages: 942 - 946
- Wang, Y., 2007, User-Centered Design for Personalized Access to Cultural Heritage. In Proc. 11th International Conference on User Modeling (UM'07) Doctor Consortium Session, Greece, 2007.
- Wang, Y., Cena, F., Carmagnola, F., Cortassa, O., Gena, C., Stash, N. and Aroyo, L., 2008, RSS based Interoperability for User Adaptive Systems, in Proc. Adaptive Hypermedia and Adaptive Web-Based Systems (Ah'08), Germany, 2008
- Weiser, M., 1991. The computer for the 21st century. Scientific America., Sept., 1991, pp. 94-104; reprinted in IEEE Pervasive Computing, Jan.-Mar. 2002, pp. 19-25



- Wilson, C., Boe B., Sala, A., Puttaswamy, K. P. N. and Zhao, B. Y., 2009, User interactions in social networks and their implications. In EuroSys '09: Proceedings of the fourth ACM european conference on Computer systems, pages 205–218, New York, NY, USA, 2009. ACM.
- Wolens, F., 2010. Facebook Security Response:spokes man Facebook Public Policy.[online] Available at:  
<<http://www.theindychannel.com/news/25841727/detail.html>> [Accessed 24 December 2010)
- Woodruff, A., Aoki, P., Hurst, A. and Szymanski, M., 2001. Electronic Guidebooks and Visitor Attention. Proc. International Cultural Heritage Informatics Meeting 2001, Milan, Italy, pages 437-454, Sep. 2001.
- Yao, Y. Y., 1995, Measuring Retrieval Effectiveness Based on User Preference of Documents, Journal of the American Society for Information Science, vol. 46, no. 2, pp. 133 -145,1995.
- Yates R. B. and Neto B. R., 1999. Modern Information Retrieval, New York: Addison Wesley, 1999, pp. 544.(Yates and Neto, 1999)
- Yates, R. B. and Neto, B. R., 1999, Modern Information Retrieval, ACM Press/Addison Wesley, 1999.
- Yih W. T., Goodman, J., and Carvalho, V. R., 2006. Finding advertising keywords on Web pages. In WWW, 2006.
- Yu K., V. Tresp V. and Yu S., 2004. A Nonparametric Hierarchical Bayesian Framework for Information Filtering. In Proceedings of the 27th Annual International Conference on Research and Development in Information Retrieval, pp. 353-360, 2004
- Zhai, C. Z., 2008, Statistical Language Models for Information Retrieval, Synthesis Lectures on Human Language Technologies, 2008, Vol. 1, No. 1 , Pages 1-141

- Zhang, M., and Alhajj, R., 2010. Skyline queries with constraints: Integrating skyline and traditional query operators. *Data Knowledge Engineering*. 69, 1 (January 2010), 153-168.
- Ziegler, S. N., McNee, S. M., Konstan, J. A. and Lausen, G., 2005, Improving Recommendation Lists through Topic Diversification, in *Proceedings of the Fourteenth International World Wide Web Conference (WWW 2005)*, pp. 22-32.
- Zinman, A. and Donath, J., 2007, Is britney spears spam? Paper presented at the Fourth Conference on Email and Anti-Spam, 2007.
- Zuckerberg, M., 2009, Governing the Facebook Service in an Open and Transparent Way.[online] Available at:<<http://blog.facebook.com/blog.php?post=56566967130>> [Accessed 26 February 2010]



# Appendix A

Precision Recall evaluation tables

Cut-off point 1		R <sub>norm</sub>		Cut-off point 3		R <sub>norm</sub>		Cut-off point 6		R <sub>norm</sub>	
cheri	V&A	cheri	V&A	cheri	V&A	cheri	V&A	cheri	V&A	cheri	V&A
1	0	1	0	2	1	0.66	0.33	4	2	0.66	0.33
1	0	1	0	2	0	1	0.5	3	0	0.5	-0.5
1	1	0.5	0.5	3	3	0.5	0.5	4	6	0.6	0.8
1	1	0.5	0.5	3	3	0.5	0.5	6	6	0.75	0.75
1	0	1	0	3	0	1	0	4	1	0.7	0.1
1	1	0.5	0.5	3	2	0.6	0.4	6	2	0.87	0.37
1	1	0.5	0.5	3	3	0.5	0.5	6	4	0.8	0.6
1	1	0.5	0.5	3	2	0.6	0.33	6	4	0.8	0.6
1	1	0.5	0.5	3	2	0.6	0.33	5	5	0.7	0.7
0	1	0	1	2	2	0.5	0.5	5	2	0.78	0.36
1	1	0.5	0.5	3	3	0.5	0.5	5	4	0.72	0.61
0	1	0	1	0	3	0	1	3	5	0.5	0.75
1	0	1	0	3	0	1	0	5	0	0.9	-0.1
1	1	0.5	0.5	3	3	0.5	0.5	6	5	0.77	0.68
1	0	1	0	3	2	0.6	0.4	4	5	0.61	0.72
1	1	0.5	0.5	3	3	0.5	0.5	6	6	0.75	0.75
0	0			0	0	0.5	0.5	3	2	0.5	0.3
1	0	1	0	3	0	1	0	1	0	-1.5	-2.5
0	1	0	1	2	2	0.5	0.5	3	5	0.5	0.75
0	1	0	1	1	2	0.33	0.66	1	2	-0.16	0.17
1	0	1	0	3	1	0.75	0.25	6	5	0.77	0.68



# Appendix B

## Identifying a User's Profiles across Social Networks

The first task in hand is to identifying a user across several social networks. Identifying and relating users profiles which are scattered across Web, will enable us to gather as much information as possible about a user's interest.

Data portability in the Social networks has recently gained a lot of attention. Users shared a lot of personal data with propriety databases in order to communicate with others in the network, this data is locked within the network, which resulted in a lot of valuable information loss, that otherwise could have assisted in understanding the user better. This information lock was once considered as advantage by the networks however with the advent in social network technologies and ways of use, the thought is now questionable. Opening data to the world now means allowing developers to build new and interesting applications over it that in turn attracts more users to participate in the network and spend more time. For example Facebook applications have played a vital role in its popularity. An interesting work here is that of Google's Social Graph API. The Google's social graph API *makes information about the public connections between people on the Web, expressed by XFN and FOAF markup and other publicly declared connections, easily available.*

### ***Our Approach***

We perform the task of user identification, as a two-step process.

1. Front End login and
2. Google's Social Graph API

#### **1. Front End login:**

This collects the required information from the user to start the Identification process over the Web. It requires the user to provide his/her Webpage and blog URIs along

with some social networking sites usernames. This is all that is required to start the identification process.

A similar approach is used by the TAGora project however our approach is different from them in the following ways.

We utilise and incorporate information from data other than tags as well to enrich the system.

We utilise the Google social graph to find as many connections about a user as possible. We do not require a user to enter any of his/her passwords on our application instead we re-direct to the original site so that the user feel more secure.

For snapshots and some code details of the login process please refer to Appendix B.

## 2. Social Graph API:

Our architecture utilise the Google Social Graph API to identify different Web pages related to a person across the Web. We use the *Site Connectivity Application* to identify different Web pages that might be related to the person.


Info on your connected sites		
The Score column shows how many of your claimed sites can be reached from each url.		
Your site	Connected to	Score
<a href="http://users.ecs.soton.ac.uk/km/">users.ecs.soton.ac.uk/km/</a>		1/2
 <a href="http://ecs.soton.ac.uk/people/km">ecs.soton.ac.uk/people/km</a>		1/2
To improve the score of the less linked sites, add a rel="me" link back to your main page		
Possible connections		
Other sites that link to one of yours claiming to be you.		
Site	Connected to	
<a href="http://linkedin.com/pub/8/312/84">linkedin.com/pub/8/312/84</a>	<a href="http://users.ecs.soton.ac.uk">users.ecs.soton.ac.uk</a>	
<a href="http://linkedin.com/pub/kirk-martinez/8/312/84">linkedin.com/pub/kirk-martinez/8/312/84</a>	<a href="http://users.ecs.soton.ac.uk">users.ecs.soton.ac.uk</a>	

Figure Sample User Information from Google Connect In the figure the top sections shows the URIs that are connected to a person's Webpage. The bottom portion shows those sites that have a link to the person's site and thus are possible connections.

These URIs that are retrieved are public links on Websites marked up using open standards like XFN and FOAF, designed to express relationships online. For example in this case, in order to use XFN to connect sites, add `rel="me"` to your link like this:

```
<a href="your URI comes here" rel="me" >me</a>
```

In our case

```
<a href="http://users.ecs.soton.ac.uk/km/" class="url" rel="me"
target="_blank"> My Website</a>
<a href="http://www.glacsWeb.org" class="url" rel="me"
target="_blank"> My Website </a>
```

We utilise the “*otherme*” method in the Google Graph API that helps locate related identifiers for a person and hence can prove useful. Other techniques used to identify same user profiles are; matching user names and real name strings from profiles across different social Web sites .





# Appendix C



## CONSENT FORM (Version 1.b)

Study title: **User Evaluation of Cheri (*Cultural Heritage Semantic Browser and Recommender*) System.**

Researcher name: Salma Noor and Kirk Martinez (supervisor)

Ethics reference: E/11/04/004

### PARTICIPANT DETAILS

These will be held securely on a Southampton University password-protected server and deleted on completion of the PhD Research (October 2011 at the latest). They will be kept separate from survey data.

1. Your name:
2. Institute name:
3. Your contact (email):

*Please initial the box(es) if you agree with the statement(s):*

Yes/No

I have read and understood the information sheet (2011-04-12/version#1.b)  
and have had the opportunity to ask questions about the study

☐

I agree to take part in this research project and agree for my data to be used for the purpose of this study

☐

I understand my participation is voluntary and I may withdraw at any time without consequence

☐

I understand that I can leave blank any question which I am unwilling or unable to answer

☐

*Additional consents (not required for survey participation):*

I am willing to be contacted by email with follow-up questions.

☐

I am willing for the information provided by me to be used as an illustrative case study within a PhD thesis. I am aware that I am entitled to withdraw this consent at any time prior to submission of the thesis without my legal rights being affected.

☐

Name of participant (print name) .....

Signature of participant .....

*(this can be typed in case the form is to be emailed)*

Name of Researcher (print name): .....Salma Noor.....

Signature of Researches .....

Date.....

**Note: (In case of remote participation)**

Please fill in the form above and email from a personally identifiable email address (such as work or university) to: sn07r@soton.ac.uk

## Participant Information Sheet

Study Title: **User Evaluation of Cheri (*Cultural Heritage Semantic Browser and Recommender*) System.**

Researcher: *Salma Noor*

Ethics number:

**Please read this information carefully before deciding to take part in this research. If you are happy to participate you will be asked to sign a consent form.**

### ***What is the research about?***

My name is Salma Noor and I am conducting research for a PhD thesis on the use of Social networking sites data as context for making personalized recommendations.

This research explores the potential of utilising social-Web data as a source of contextual information for searching and information retrieval tasks. While using a semantic and ontological approach to do so, it works towards a support system for providing adaptive and personalized recommendation of Cultural Heritage Resources.

This study outlines an evaluation that is carried out to demonstrate the scrutability (user-acceptance and accuracy) and the dynamic and adaptive nature of the feedback mechanism in Cheri which is a prototype cultural heritage recommender system.

It is intended that the results of this evaluation will provide a useful baseline for the use of social networking data to specifically address the cultural heritage related needs of a general Web user. This research is being paid for by the Higher Education Commission of Pakistan under the Faculty Development program for Frontier Women University Peshawar.

### ***Why have I been chosen?***

As a user of the social networking sites such as facebook, you are able to provide important perspectives on the merits of the use of social network data to gather user

interests. And the potential in such data for providing personalised access to Web resources.

***What will happen to me if I take part?***

If you consent to take part you will be asked to provide some basic details about yourself in the consent form. The details of the information asked and how it is protected is as follows.

*Participants Details (asked in the consent form):* These will be held securely on a University of Southampton password-protected system and deleted on completion of the PhD Research (October 2011 at the latest). They will be kept separate from survey/study data. The following data is collected in consent form:

- Your name
- Institute name
- Contact (email)

You will then be asked to visit the link: <http://degas.ecs.soton.ac.uk/cheri/cheri-v1.5>

To evaluate the system and will be provided with a questionnaire to answer alongside.

*Data to be Collected During this Study:*

At the beginning of the study the system requires the user to login to their facebook account and allow the extraction of their interest information that includes the following:

- Explicitly mentioned interests in user's facebook profile under the heading Arts and entertainment (Music, Books, TV, Movies) and Activities and Interests.
- Geo coordinates (user location at the time of use of the system)

The user is explicitly asked beforehand if they want the system to proceed and extract this data from their online facebook profile and is given the opportunity to withdraw if he/she does not wish to do so.

The extracted information will be held securely on a University of Southampton password-protected server, and is kept separate from any of the Participants details (obtained in the consent form) in order to retain anonymity. Furthermore, the location

information is kept separate from the interest information of the user to ensure user anonymity.

*Data Collected in Questionnaires:*

For Questions asked during the study kindly see the Attached Questionnaires. The questions asked are mostly about the participant's views on how the system responds to the different tasks and does not involve any personal information on their part.

**Subject to your additional consent**, I may contact you by email with some follow-up questions.

*Are there any benefits in my taking part?*

There are no personal advantages (inducements) to be gained from taking part, but the overall results of this study will contribute towards a better understanding of user perspective in adoption and potential use of such technologies in future.

*Are there any risks involved?*

There are no personal risks involved to our knowledge.

*Will my participation be confidential?*

Any personally identifying information provided by you will be held in accordance with the Data Protection Act and University policy on a password-protected computer at the University of Southampton. It will be deleted on completion of my PhD (October 2011 at the latest). Association of the details with the survey data will be by means of an identifying codename. All possible measures will be taken to ensure anonymity in publication. Under the terms of the Data Protection Act you can request a full copy of all information held about you as a result of this survey by contacting the ECS School Office and citing the above ethics number: Electronics and Computer Science, University of Southampton, SO17 1BJ, United Kingdom. Email: [school@ecs.soton.ac.uk](mailto:school@ecs.soton.ac.uk).

***What happens if I change my mind?***

You are able to withdraw or amend your data at any time and for any reason prior to submission of the thesis without your legal rights being affected in any way. Should you request this, both personal and survey data provided by you will be deleted.

***What happens if something goes wrong?***

In the unlikely case of concern or complaint, you are welcome to contact:

Lester Gilbert, Chair of Southampton University ECS Ethics Committee:

[l.h.gilbert@soton.ac.uk](mailto:l.h.gilbert@soton.ac.uk)

***Where can I get more information?***

The participant can contact the Investigator at any time during or after the course of study to ask any further questions.

Investigators contact:

*Name:* Salma Noor,

*Email:* [sn07r@ecs.soton.ac.uk](mailto:sn07r@ecs.soton.ac.uk)

## Pre-Evaluation Questionnaire

1. Do you browse the internet while you are on vacation abroad for interesting places to visit?

☐ Yes ☐ No

If yes where do you search? (e.g., Google)?

Comment.....

2. How often do you visit a museum Website or look for information online before visiting a museum.

Never ☐ Very Rarely ☐ Sometimes ☐ Very Often ☐ Always ☐

3. Do you search for cultural heritage related information online?

Never ☐ Very Rare ☐ Sometimes ☐ Very Often ☐ Always ☐

4. Have you ever used a handheld travel guide system?

Yes ☐ No ☐

If “Yes” for question 4 go to 5 else go to 6

5. How has your experience been with such a search system?

Excellent ☐ Good ☐ Average ☐ Not very useful ☐ frustrating ☐

6. Would you be interested in a personalised, tailored to your interest, art recommending facility?

Yes ☐ No ☐

7. Would you use such a facility if available? If no why not?

Yes ☐ No ☐

Comment .....

8. Do you understand personalization?

Yes ☐ No ☐

If “Yes” go to 9 else proceed to next section (i.e. evaluation questionnaire)

9. What do you mean by personalization (describe in your own words)?

.....  
.....

10. Is personalization important for you?



Yes ☐ No ☐

11. Is personalization a good thing in your point of view?

Yes ☐ No ☐

If No why not? .....

## Evaluation Questionnaire

### Task 1:

- Open the cheri Web link. (<http://degas.ecs.soton.ac.uk/cheri/cheri-v1.5/index.php>)
- Follow the instructions to connect to your facebook account
- Authorize the transfer of your interest information.
- View the generated interest page and answer the Questions 12 to 14 below.

12. Were your interests correctly transferred from your facebook profile to the Cheri system?

Yes ☐ No ☐

13. Did you face any problems during the transfer?

Yes ☐ No ☐


If yes please specify?

Comment .....

14. Were the suggested links displayed correctly?

Yes ☐ No ☐

### Task 2:

- Read the interest links generated against each of your acquired interests and see if they point to the right concept and right category.
- Check the wiki link  given in front of it and see if they point to the same concept as mentioned in your facebook profile.
- Answer the Question number 15 below.

15. How many of the interest pointed to the concept you meant in your facebook profile?

Number / Total (No of concepts in your profile)

### Task 3:

- Open the tab marked **“Web Recommendations”**
- If you see a **“do you mean”** field in it **that means the system was unable to identify the right concept for your interest.**
- In this case see if you can identify, if the right concept is listed in the links suggested below the **“do you mean”** field.
- Answer questions 16-18 below.

16. How many of the interests presented to you had the field **“do you mean”** in them?

Number / (Total number of concepts in your profile)

17. How many interests that had a **“do you mean”** field had the right concept mentioned as the suggested links?

Number / (Total number of concepts in your profile)

18. How many concepts were you unable to resolve through the Cheri system?

Number / (Total number of concepts in your profile)

- Check the Website links suggested to you based on your interests gathered and see if you find them relevant. Then answer question number 19

19. Were the Web links relevant to your interest?

Yes ☐

No ☐

- Check the Web images suggested to you and see if they are relevant. Then answer question number 20.

20. Were the Web images relevant to your interest?

Yes ☐

No ☐

- About task number 3 (Web recommendations).

21. Was the Web recommendation option useful? On a scale of 1 to 5

• Yes

• No

• Number

22. What was your satisfaction level? On a scale of 1 to 5

Number

**Task 4:**

- See the tab marked “**V&A Recommendations**”
- Check if the artwork (image) relates to your interest it is listed under.
- Choose an image or two you are interested in.
- There are some selectable properties of the image in front of it (artist, object, place and time). Select one or more properties to get future suggestions on art work. See if they are to your liking.
- Answer Questions 23-28 below

23. Was the selection mechanism for modifying the search according to the artwork properties, easy to use?

Yes ☐ No ☐

24. Were the results relevant to your feedback provided to the system through selection of properties?

Yes ☐ No ☐

25. Did you find the feedback (through selection of properties) mechanism useful?

Yes ☐ No ☐

26. Would you have preferred any other feedback mechanism? If Yes please state what other mechanism?

Yes ☐ No ☐

Comment .....

27. Did you face any issues in using the cheri feedback system? If Yes please explain?

Yes ☐ No ☐

Comment .....

28. What do you suggest can be done to improve the feedback system?

------------------

**Task 5:**

- Click the tab named “**place of origin**” on the top of the page
- This gives you a global map view of the artwork recommended to you on the bases of your interests captured.
- Answer the Question number 29 and 30.

29. Did you face any issues in using the map based representation of the artwork? If “Yes” please explain?

Yes ☐ No ☐

Comment .....

30. What do you suggest can be done to improve the view?

**Task 6:**

- Click the tab “**Near you**” and click the button shown on the newly displayed page.
- A map displaying Cheri’s personalized Geo-Based recommendations appear.
- Scroll over the artwork-thumbnails on the map and see in the popup details if the artwork suggested on the map is from your current location.
- Answer question number 31 and 32

31. Did the system register your current location i.e., is the map centred at your current location (e.g., Southampton)?

Yes ☐ No ☐

32. Are the results presented as thumbnails over the map relevant to your current location?

Mention in number how many of the total are unrelated (if any) .....

- Are there any red marked links on the map (these are the artefacts that match both your interest and your current location)

33. How many of the results out of total have red markers attached to them?

Yes ☐ No ☐

34. Did you face any issues in using the cheri Geo based recommendation viewer? If Yes please explain?

Yes ☐ No ☐

Comment .....

35. What do you suggest can be done to improve the system?

**System Benefits and issues:**

36. How do you believe you can benefit from this system?

37. What issues if any have you faced in using the cheri system?

## Post-Evaluation Questionnaire

38. How much is privacy in social networks important for you?

Never ☐ Very Rarely ☐ Sometimes ☐ Very Often ☐ Always ☐

39. Are you happy to share your information with other users of the social network?

Never ☐ Very Rarely ☐ Sometimes ☐ Very Often ☐ Always ☐

40. What information will you willingly share on Web with the following 3 categories of people tick as appropriate?

**1) With friends:**

Location (country/city) ☐ Interests ☐ Hobbies & Activities ☐

Profession Info ☐ Status Updates ☐

**2) With everyone (in your social networking circle i.e. people you have added to your connections but are not necessarily your friends):**

Location (country/city) ☐ Interests ☐ Hobbies & Activities ☐

Profession Info ☐ Status Updates ☐

**3) With strangers (who are not connected to you but are using the same social networking platform e.g. facebook users):**

Location (country/city) ☐ Interests ☐ Hobbies & Activities ☐

Profession Info ☐ Status Updates ☐

41. Are you comfortable in adding people, you do not know in real life as your facebook friends?

Yes ☐ No ☐

42. Do you accept an 'add as your friend' request on facebook without knowing the person in real life?

Yes ☐

No ☐

If yes what is your acceptance criteria (e.g., the person belongs to the same workplace as you, or is from your hometown) please comment

Comment.....

43. How many of your facebook friend you do not know from your real life (i.e. have met in person)?

Approximately (in percentage) .....

Thank you for taking the time to fill this questionnaire, your help is much appreciated ☺

