University of Southampton

Faculty of Engineering, Science and Mathematics

School of Electronics and Computer Science

Tag Disambiguation Based on Social Network Information

by

Syed Sumair Qasim

23rd September, 2011

Project Supervisor: Dr Mark J Weal

Second Examiner: Dr David E Millard

A dissertation submitted in partial fulfilment of the degree of

MSc Software Engineering

by examination and dissertation

Abstract

Within 20 years the Web has grown from a tool for scientists at CERN into a global information space. While returning to its roots as a read/write tool, its entering a more social and participatory phase. Hence a new, improved version called the Social Web where users are responsible for generating and sharing content on the global information space, they are also accountable for replicating the information. This collaborative activity can be observed in two of the most widely practised Social Web services such as social network sites and social tagging systems. Users annotate their interests and inclinations with free form keywords while they share them with their social connections. Although these keywords (tag) assist information organization and retrieval, they suffer from polysemy.

In this study we employ the effectiveness of social network sites to address the issue of ambiguity in social tagging. Moreover, we also propose that homophily in social network sites can be a useful aspect is disambiguating tags. We have extracted the 'Likes' of 20 Facebook users and employ them in disambiguation tags on Flickr. Classifiers are generated on the retrieved clusters from Flickr using K-Nearest-Neighbour algorithm and then their degree of similarity is calculated with user keywords. As tag disambiguation techniques lack gold standards for evaluation, we asked the users to indicate the contexts and used them as ground truth while examining the results. We analyse the performance of our approach by quantitative methods and report successful results. Our proposed method is able classify images with an accuracy of 6 out of 10 (on average). Qualitative analysis reveal some factors that affect the findings, and if addressed can produce more precise results.

Acknowledgements

I would take this opportunity to express my gratitude to my supervisor Dr Mark J Weal for his invaluable support and guidance. His supervision helped me throughout the stages of dissertation. Besides that I would also like to thank the group supervisors Dr David E Millard and Dr Thanassis Tiropanis for continued support and advice in successfully completing the dissertation. I am also grateful to the ECS help desk for providing assistance in the implementation of the experiment.

Lastly but in no way the least, I would like to acknowledge the continuous support and motivation provided by my family and friends which kept me determined while writing my MSc dissertation.

Table of Contents

List of Figures

List of Tables

1 Introduction

1.1 Project Scope

In today's world of Web 2.0 where users are responsible for generating and sharing content on the global information space, they are also accountable for replicating the information. This particular case can be observed in social tagging or collaborative tagging systems which allow users to make use of keywords(tags) to describe publically available Web content and this entire exercise constitutes to a concept called Folksonomy as described by (Vanderwal, 2007). Features like these added extra values to the social networking services and the network effect of some of these services attracted millions of users and are still growing to constitute what is called the social era of the Web. One of the major reasons behind the success of these online networking services is the fact that they employ numerous multidisciplinary concepts from social sciences, psychology, information retrieval and knowledge organization. They provide simple, inexpensive ways to organize members, arrange meetings, spread information, and gauge opinion.

Besides being a medium for disseminating information, the Web in recent years has also become an important platform of social interactions. Wikipedia[1] represents an excellent example of how collaborative Web users have become. Facebook[2], MySpace[3], Renren[4], Orkut[5] to name a few social networking sites allowing users to share their interests and to keep track of what their friends are doing. All this contribution is in the form of freely-chosen descriptive keywords that are used by Web users to describe, organise, share and retrieve Web resources. Some obvious examples of these keyword driven collaborative systems are Delicious[6] and Flickr[7]. Golder and Huberman (2006) analyse the structure of collaborative tagging systems. One of the problems they identify is of polysemy, words having multiple meanings. Polysemous keywords greatly hamper information retrieval. They dilute the Web search results by returning results that are related but not applicable as the context is not defined.

The motivation of this study lies in evaluating the strength of one social Web service to address the weakness of the other. There has been number of studies undertaken in applied sciences to resolve the issue of tag disambiguation. In computational linguistics, this is termed as word-sense disambiguation (WSD) (Ide & Jean, 1997) which is an open problem of natural language processing that deals with the process of identifying different senses when the word has multiple meanings (polysemy). Using word sense disambiguation has shown improvement in information retrieval and hypertext navigation. Many supervised and unsupervised methods have been developed, however unsupervised learning is still a challenge for WSD researchers. The fundamental assumption is that similar senses occur in similar contexts, and thus senses can be induced from text by clustering word occurrences using some measure of similarity of context(Palo & Alto,1998), a task referred to as word sense induction or discrimination.

[1] http://www.wikipedia.org/
[2] http://www.facebook.com/
[3] http://www.myspace.com/
[4] http://www.renren.com/
[5] http://www.orkut.com/
[6] http://www.delicious.org/
[7] http://www.flickr.com/

Several studies have employed different approaches to get hold around the contextualisation of tags which includes applying data mining techniques on the large data set to form clusters. Yeung et al (2009) researched on the associations of a single tag with other tags, users and documents in a folksonomy. However there is little work done on resolving tag meaning with respect to social networks which has been the most widely used service of the Web 2.0. Therefore, we propose that by making use of information available in social networks we can contextualise the keyword/tag with more precision as compared to the techniques which don't use the social context. We believe that this would also help improve interest matching in social networks. The novelty of this study is in employing a sociology concept called Homophily (McPherson et al., 2001) (i.e., "love of the same") and proposing that social network information of a user and friends of user can improve tag disambiguation in a Web browsing session, particularly when querying to an image search engine.

1.2 Research Objectives

This dissertation focuses around the area of tag disambiguation on the Social Web with emphasis on social networks. The problem statement that we plan to study is:

"The effectiveness of social network information in resolving the problem of ambiguity in social tagging"

By conducting the research work described in this study, we aim at answering the following research questions:

- Can social network information be used for tag disambiguation?
- Can homophily in social networks help improve the effectiveness in disambiguating the polysemous tags?

The two questions are interrelated to each other. Firstly we want to know whether the social network information is a worthwhile data source for tag meaning resolution. This activity involves data collection from a social network and applying a state of the art algorithm for resolving tag ambiguity. Secondly, we want to investigate the homphilous nature of social networks. For example, we can use the findings of our initial experiment and modify it to allow friend's social network information. Can these approaches produce better result? How do we evaluate our methods and their outcomes? Does our method improve information retrieval? We will require a better understanding of relevant literature to answer these questions and it will also allow us to gain a better understanding of the Social Web, eventually improving Social Web applications to facilitate better user interactions on the Web. In practise, we would like to test the following hypothesis concerning polysemous nature of tags:

1.2.1 Hypothesis 1

User's collaborative nature is specially taken into consideration when modelling associations in a Social Web application; likewise such shared information also reflects user's interests and therefore can be employed to disambiguate tags.

1.2.2 Hypothesis 2

Homophily is one of the evident characteristics in a social network. This similarity in interests can be put to use in a disambiguation framework to enhance information retrieval from folksonomies.

1.2.3 Potential Challenges

As this study involves empirical studies on social networks, word sense disambiguation and information retrieval, the methodology should include some means of data collection, user collaboration, testing environment and assessment measures. The initially challenge is of information extraction from a social network, cleaning it and preparing it for the experiment. We then have to implement the disambiguation technique (clustering/classification algorithm), for this we might use or refactor one of the existing algorithms. In depth knowledge of existing methods is required. Finally our last milestone would be identifying an evaluation criterion that could formalise and summarise our findings.

1.3 Report Structure

The first few chapters of the report would comprise of sections focusing on the introduction and background of the areas related to the topic. These sections will present all necessary information about the project, relevant literature and most recent methodologies undertaken to solve disambiguation problem. Moreover these approaches will be critically analysed in order to set out the methodology chosen for this study. The next 2 chapters will have a detailed description of the steps we took for this study including information extraction, storage, implementation and testing. We are going to evaluate our results in chapter 5 of the report; this will include our examination of results, analysis and discussion on our findings. Finally in chapter 6 we will conclude and summarise our approach and the findings complying with our hypotheses.

2 Background

In order to have a more in-depth understanding of concepts surrounding the problem we have to analyse and critique the work done in the related areas. In order to do that we have set out a literature map that will be useful to navigate across the vast literature gathering necessary and up to date information. We will go about reviewing the literature around social networks w.r.t the following areas of applied and social sciences:

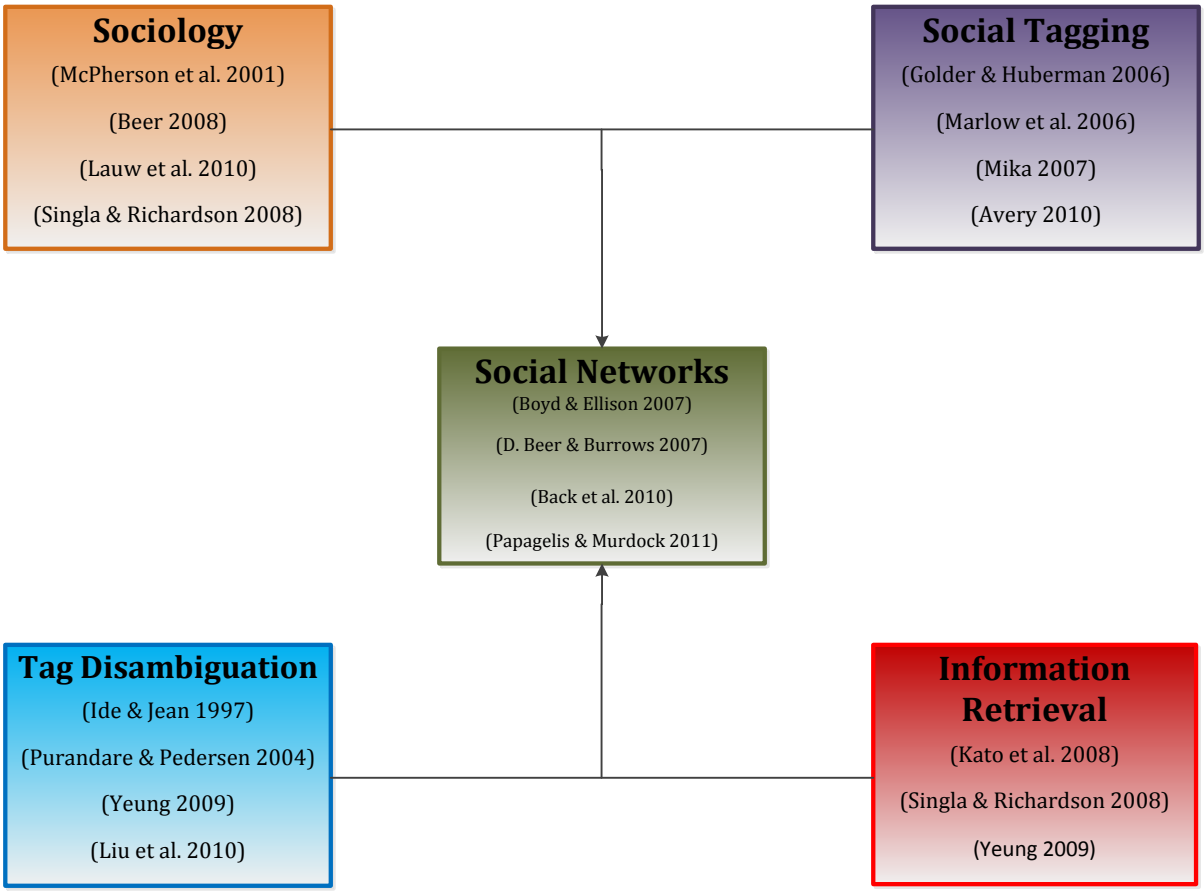

Figure 2-1: Literature Map

2.1 The Advent of Social Web

Within 20 years the Web has grown from a tool for scientists at CERN into a global information space. While returning to its roots as a read/write tool, its entering a more social and participatory phase. Hence a new, improved version called the Web 2.0. However for semantic Web scientists, the essence lies in the transformation to an information space in which machine-readable data, infused with some sense of 'meaning', the semantic Web. Of course this transformation can only take place with the help of user contribution which has been happening for a while. Numerous applications such as Wikipedia, Flickr, Facebook and YouTube etc. have overruled the perception of many mavens who underrated people's desire to use the Web to socially mediate their information environments and communications. Chi et al refer this as the Social Web(Chi & Alto, 2008) where people use the Web 2.0 services to fulfil their social needs such as information retrieval, sharing and bookmarking, and collaboration. The following figure calibrates the degree of collaboration of these Web 2.0 services.

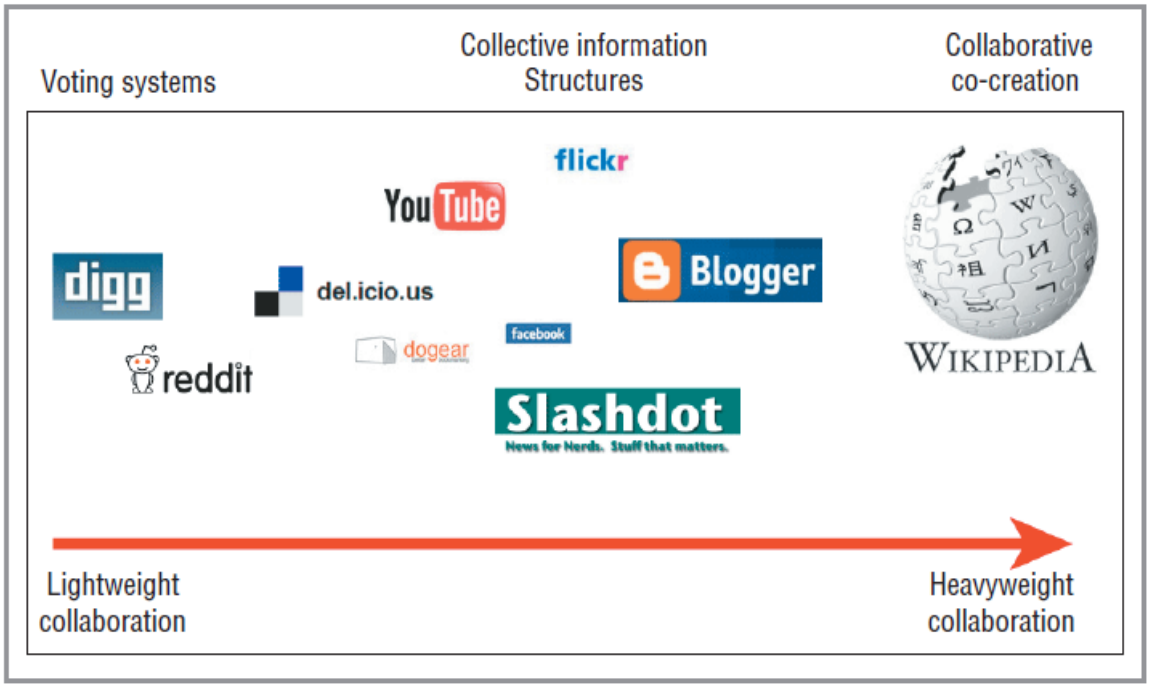

Figure 2-2: Web 2.0 Services and Collective Intelligence

Two of the most used services of the Social Web are social network sites and collaborative tagging systems. These two services are the natural complements of each other. One caters the need of information propagation, friendship and extroversion, the other deals with information organisation and retrieval. Collectively they facilitate the modern day Web user to maintain a social and structured portfolio on the worldwide information space.

2.2 Social Networks

Social networking sites (SNS) are perhaps the most socially accepted of the Web 2.0 applications, particularly as the number of users continues to grow and as they integrate a range of other Web 2.0 applications (mashup). In sociology, a social network is a social structure made up of nodes which are connected by one or more specific types of interdependency. These type(s) can be friendship, kinship, common interest, financial exchange, sexual relationships, or beliefs. Beer and Burrows (2007) explain the sociological process of cultural transition occurring through the induction of Web 2.0 services. Among these, there are three interrelated issues that require sociological engagement: the changing relations between the production and consumption of content; the mainstreaming of private information posted to the public domain; and, the emergence of a new rhetoric of 'democratisation'. The authors claim that Web 2.0 services particularly social networking sites have completely changed the way people socialise, it has introduced the notion of Web identity and other linked concepts like influence, power and trust have also got a new meaning. Social networking sites have become an integral part of the daily lives of millions of users due to the very fact that 'Man is a social animal'. These sites generate huge network effect because they have engaged primitive social psychology phenomena. The Big-Five framework suggests that most individual differences in human personality can be categorized into five broad domains. Extraversion is one of those personality traits that is quite visible in social networking sites. Gosling et al (2003) assert that extraversion is the best of these traits in personality judgement. Several other studies conducted on social networking site have identified the presence of social network concepts. (A., Orkut, & Eytan, 2003) present the

analysis of an online community at Stanford University and report the occurrence of phenomena such as the small world effect, clustering and the strength of weak ties. Thus, social networking sites do exhibit the conventional characteristics of any social group and also allow us to characterize social ties and identify what factors influence friendships. Though these characteristics are not new, there application has been proved in a more ubiquitous way, for instance (Milgram, 1967) explains the small world phenomena as a social network where every person can be reached within six steps (six degree of separation). Milgram's work is considered as the benchmark in social sciences. Many researchers have carried on his work in order to argue and verify his claims in accordance with changing needs of technology. Rosen (2007) explains that this small world experiment was conducted by Duncan J. Watts a professor at Columbia University, he used email instead of letters and it was not restricted to United States as in the original study. Results suggest that Milgram might have been right, as messages reached their destination in five to seven steps on average. Social networking theorists equally support the smallness of our wireless world. Albert-László Barabási asserts:

> "*The world is shrinking because social links that would have died out a hundred years ago are kept alive and can be easily activated. The number of social links an individual can actively maintain has increased dramatically, bringing down the degrees of separation*"
>
> (Barabási 2002)

Furthermore he claims that in this world of Social Web the steps could be reduced to just three.

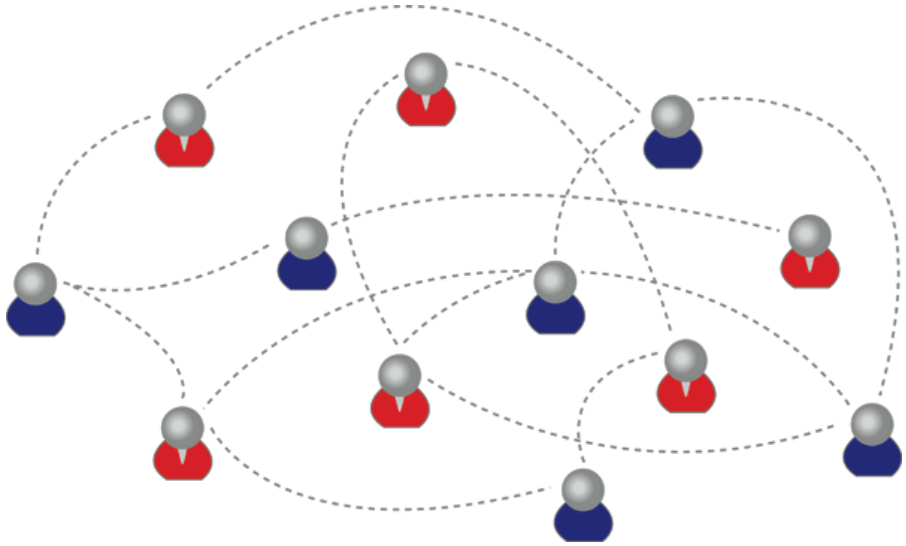

Figure 2-3: Social network

Today's social networking sites organize themselves with analogy of the person, who has a personal profile that includes hobbies, interests and relations with other profiles. Consequently, one's introduction into this world is through the disclosure of personal information. Boyd and Ellison define social networking site in a more formal manner:

> *"Social network sites 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 & Ellison, 2007)

Their study presents an in-depth perspective on the history of such sites discussing key changes. They argue that the growth of these social enterprises points to a shift in the organization of online communities. While websites dedicated to communities of interest still exist and prosper, SNSs are primarily organized around people, not interests. This also supports the concept of unmediated social structures, where ''the world is composed of networks, not groups'' (Wellman 1988). Thus with the introduction of such social Web services a new organizational framework for online communities is introduced. Moreover these sites have great research potential as the two researchers highlight by showing how these networked practices reflect and amend day to day behaviour especially how people present themselves online. All of this activity offers great research opportunities to social Web scientists resulting in scholarships that help explain these online and offline social behaviours. Boyd and Ellison have become prominent names in this field as they operate at the state of the art of what is going on.

Dr David Beer (2008) in his response to Boyd and Ellison critique the definition and theory supporting their article. He argues the use of term 'network' just because it broadens the scope and decreases the emphasis on relationship initiation in SNS. Although Boyd and Ellison's definition also accommodate services like blogs, wikis and folksonomies, Beer suggests the use of more general term like Web 2.0 could act like an umbrella to all these services including social network(ing) sites. There is a great deal of overlap among these services as they share aspects like tagging, profile, collaboration and friending. The element of 'networking' is present in almost every service and this also provisions the introduction of mashups into the mainstream Web applications. Furthermore, he urges that there is more to SNS than just analysing user behaviour and harvesting collective intelligence. There is a need to handle challenges like the growing 'rhetoric of democratisation' that has emerged and ushered in Web 2.0 (Beer & Burrows, 2007). Other issues that need attention are capitalist interests, data usage by third party, the organising power of algorithms, privacy issues, social identity, influence and how SNS can be understood as collections of transactional data about a vast population of users. These are the research areas that are under study and continue to make the user experience more social yet collaborative on the Web.

2.2.1 Social Networks and User Behaviour

With the inclusion of SNS in the civil society, many factors like influence, power and trust are now visible through user activity on the Web. These SNS enable users to connect with each other, share and find content, and propagate information. Some of these sites provide social links (e.g., LinkedIn, Facebook, and MySpace) and some provide networks for sharing content (e.g., Flickr, YouTube).

The understanding of how users behave when they connect to these sites is important in a number of ways. Performance of existing systems can be evaluated by studying user behaviour. Better models of user activities in SNSs are vital in social studies as well as in viral marketing which ultimately help revenue models. Moreover, analysing these Web browsing sessions also help in designing the next-generation internet infrastructure and content distribution systems. By

mentioning the prominence of user activity particularly on social networks, one can question the correctness of this contribution and the influence, and correlation observed between the actions of user and the number of friends in the network. Studies show that there has been work done on answering these questions.

Researchers have gathered data from renowned SNS and performed theoretical and empirical experiments on validating the accuracy of user contribution. Facebook is one of the most prominent names in SNS with more than 750 million users and an average user has 130 friends[8]. It is also the source of data and numerous research ventures. Investigation conducted on Facebook profiles to get personality impressions in accordance with the big five framework shows strong consensus on extraversion amongst other traits (S D Gosling et al. 2007), thus the data in this study conforms to the fact that the social network sites are a viable and valid source of communicating personality. Another recent piece of work validates this argument where research was conducted on 236 SNS users from US (Facebook) and Germany (StudiViz[9]). Ideal and observed ratings when calibrated against big five personality traits showed no evidence of self-idealization, hence establishing that Facebook profiles reflect actual personality not self-idealization (Back et al. 2010). Studies also show how a user's activity correlates to user's friends' social affiliation. Marlow et al. (2006) examine the tags in Flickr by a user and those placed by friends of the user. They report a correlation between social connectivity and tag vocabulary. Several other studies have statically shown the existence of social influence and correlation by proposing methods and frameworks. Aris, Kumar, and Mahdian (2008) describe the various models of correlation and perform numerical analysis(shuffle test) on the actual data from Flickr with a view to identify and measure social influence as a source of correlation between the actions of socially active users. Papagelis and Murdock (2011) have extended their work and proposed a method for detecting social influence in a social system and also highlighted the relation between influence and user credibility. However we would focus on the work of Aris et al. as they also discuss the causes of this behaviour (correlation). They assert homophily is one of the prime reasons of correlation between users in a social network. Individuals often make friends with people who are similar in interests. This phenomenon does not only exist in the traditional social networks it is also empirically evident in social networking sites as recent studies have shown.

2.2.2 Homophily In Social Networks

The common saying, "birds of a feather flock together," represents the elementary definition of homophily. The concept of homophily is not new as sociologists Lazarsfeld and Merton (1954) describe in their original formulation of homophily. It has played a vital role in understanding the dynamics of social behaviour of individuals in a network. Extensive studies have been conducted in social sciences to detect the influence of homophily from an individual and communal perspective. In a social network where individuals (nodes) are connected to other nodes, homophily is one of the factors that breeds this connection.

One of the most thorough and well cited works on explaining homophily, its causes and effects is by McPherson et al (2001). They explain that people have can be classified in different characteristics and dimensions such as genders, ethnicities, ages, backgrounds and educational qualifications etc.

[8] Facebook Statistics-https://www.facebook.com/press/info.php?statistics
[9] German Social Network- http://www.studivz.net/

Once categorised in one of these dimensions, people tend to display certain qualities common throughout the group. For instance, educated people tend to be tranquil and tolerant. This organization also limits their social connections to those with whom they share that dimension resulting in the quality to be localized in socio-demographic space. Such an activity or a tendency is called 'homophily' where interaction between similar people occurs at a higher rate than among dissimilar people hence forming networks of people(with some common agenda) . Any cultural, behavioural, genetic or material information flowing in these homophilous networks will tend to be localised. Homophily implies number of aspects concerning a social network:

1. Localization of social characteristics translates network distance into the number of connections (relationship) through which the information has to flow to connect the two nodes.
2. Any social structure that depends on these homophilous networks for information diffusion will tend to be localised in the social space.
3. As localization of social characteristics (homophily) is the key to the operation of many social structures, it can be used as an organizing concept.

Two types of homophily have been identified in the literature:

2.2.2.1 Status Homophily:

The similarity based on informal, formal or social eminence can be classified as status homophily. It includes socio-demographic dimensions such as race, ethnicity, gender, or age, and attained features like religion, education, occupation, or behaviour patterns.

2.2.2.2 Value Homophily:

The similarity based on values, beliefs and attitudes is referred to as value homophily. It includes the wide variety of activities performed to shape our future behaviour. Moreover, it emphasizes on grouping different school of thoughts without taking into account their demographics. This gives value homophily a sense of global organization

2.2.2.3 Evidence about Homophily

McPherson at al. (2001) confirm the existence of homophily in both the types; status (age, gender, race) and value or behavioural homophily. According to the literature they refer, the concept of homophily is remarkably evident across various social ties and dimensions of homogeneity which characterizes network systems as well personal networks. For instance in multicultural societies ethnicity create the simplest divides. Traits like gender, age, religion, and education also strongly assemble our relations with others. Profession, network position, behaviours, and intrapersonal values also show significant homophily, but they seem to be more specific to certain types of networks (e.g. community of Web scientists supporting open Web standards). They further discuss the baseline patterns that strongly influence networks to encourage connections within large organizations as well as smaller social spaces. Therefore we can imply:

1. Homophily exist in social networks whether demographic or behavioural
2. The similarity principle tends to localize the common trait or behaviour across the network
3. Homophily also tends to influence the future social encounters of an individual as well as a community.

2.2.2.4 Homophily in Digital World

Sociology has studied homophily in the physical world as described in the former section where geographic proximity, family ties, and organizational factors, such as school and profession play an important role in generating homophilous networks. However, these factors become less significant while studying value homophily in the digital world. Interest and thoughts are more vital in establishing homophily on the social Web. Lauw et al. (2010) study homophily using data from LiveJournal[10] and claim that relationships and interests are strongly interlinked, having common interests friendship become more probable. Similarly friends also are more likely to share common interests individually and across communities (see Figure 2-4)

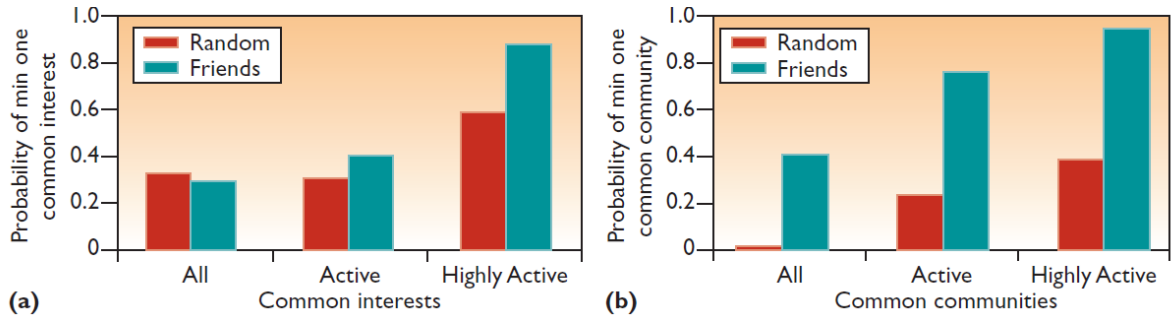

Figure 2-4: Probability of commonality for different activity levels; (a) at least one common interest and (b) at least one common community (Lauw et al. 2010).

They also suggest to study homophily using data from other social networks such as Facebook or Orkut. Doing so will allow to analyse the structural differences between the networks and how those might affect the role of interests in friendship. Another study uses Twitter[11] data and examine the relationship between homophily along diverse user characteristics and the information diffusion process on social media (De Choudhury et al. 2010). Results show that the particular attribute that can best represent diffusion depends upon the diffusion metric as well as the topic under deliberation. Hence attribute homophily plays a significant role in quantifying diffusion characteristics. Information propagation in social networks also affects the individual behaviour on the Web. The authors of a study (Singla & Richardson 2008) assert that there is a correlation between social networks and personal behaviour on the Web. They examine chat sessions of a prominent instant messaging service and reveal that people who chat with each other are most likely to have shared interests resulting in their Web searches to be the same or similar topic wise.

The findings of another study exhibit that there is a varied range of homophily in MySpace[12] in the sense of active connections tending to be more similar than would be expected if friends were chosen at random from MySpace friends (Thelwall, 2009). Significant evidence of homophily is found to be existing in sources like ethnicity, age, religion, and other offline traits (sexual orientation, country, marital status) . Moreover gender is not found to be a source of homophily as also reported by McPherson et al. Thus we can say that homophily is an evident social trait observed on the social Web while disseminating information, improving friendship, influencing personal behaviour and improving information organization.

[10] http://www.livejournal.com
[11] http://twitter.com
[12] http://www.myspace.com

2.3 Social Tagging

As we mentioned in the previous section, a very prominent feature of the Social Web is enabling users to annotate and rate online content in many social Web enterprises. Tagging is a user activity which comprises of assigning freely-chosen descriptive keywords to online content. The foremost use of keywords or tags is in information organisation, storage and retrieval on the Web. The idea of tagging was initiated by social bookmarking sites such as Delicious in 2005, since then it has gained acceptance leading into classification and sharing of images (e.g. Flickr), books (e.g. LibraryThing), and academic references (e.g. Bibsonomy and Connotea). Similarly Amazon started this feature in 2006 where it allowed users to assign tags to books and other products sold on their website (Iskold 2007). It is also used in indexing and retrieval purposes on video sharing sites such as YouTube. A report state that approximately 28% of internet users in America have used different form of tagging activities(Rainie, 2007).

Tagging has also gained importance on a number of social websites. The SNS allow tagging by enabling users to easily annotate the content (web page, photo, video etc) they publish and share by using tags to make them searchable and discoverable in future by others which ultimately gives a social aspect to tagging. (Golder & Huberman 2006) identify several roles that tags can perform for users, from topic definition to opinion forming and even self-reference. (Marlow et al. 2006) gives another interpretation of social interpretation of tags for example 'Seen Live'. (Breslin et al. 2009) in their book 'Social Semantic Web' describe the importance of tagging on the social Web particularly SNS. One of the most important features of social networking sites is the utility to upload and share content with user's network. People share, interact and socialise due to common interests related to particular objects. The object can be a movie, celebrity, technology or even a place shared with whoever is subscribed to it or just within a community. Facebook uses this feature in the form of 'Likes'. In fact they announced in 2010 these 'Likes' are a form of 'social links' , better than a link because it's related to a specific user (Cashmore, 2010). These 'Likes' very keenly capture the idea of user's interest and social behaviour as these keywords can point to any form on multimedia and social group within the network. Theoretically Peter Mika describes this activity as tripartite graph between a user, a resource and a tag (Mika 2007). (Vanderwal 2007) terms this graph as a folksonomy, a portmanteau of 'folks' and 'taxonomy'. Hence, a folksonomy is a social, collectively generated, open ended, evolving and user driven categorisation scheme. A recent piece of work studies folksonomy, its contemporary practises and how information professionals are reacting to these developments(Avery, 2010).

At present folksonomies are primarily used in social networking sites, such as Facebook. Museums, libraries, educational and corporate environments are accepting this concept now. However social tagging faces scepticism by people in the information sciences who argue that these schemes are not philosophically valid and will lead to a system breakdown (Peterson, 2006). Other information professionals agree with the potential characteristics of folksonomies enabling creative and dynamic information organization (Guy & Tonkin, 2006).

2.3.1 Ambiguity in Tags

Polysemous and ambiguous tags are inevitably common in folksonomies due to lack of semantics or interpretation for machine that reads them. In contrast to a search engine a person can distinguish between the contexts whether the tag 'apple' is used in relation to a laptop or about a picture fruit. Results from the both the contexts will be displayed on a search engine and its user's job to sort out

which one is relevant and which one is not. This activity can be inefficient depending on the number of returned search results. For example, following figure show results of images tagged as 'Apple', as it is visible that the results contain images from multiple contexts.

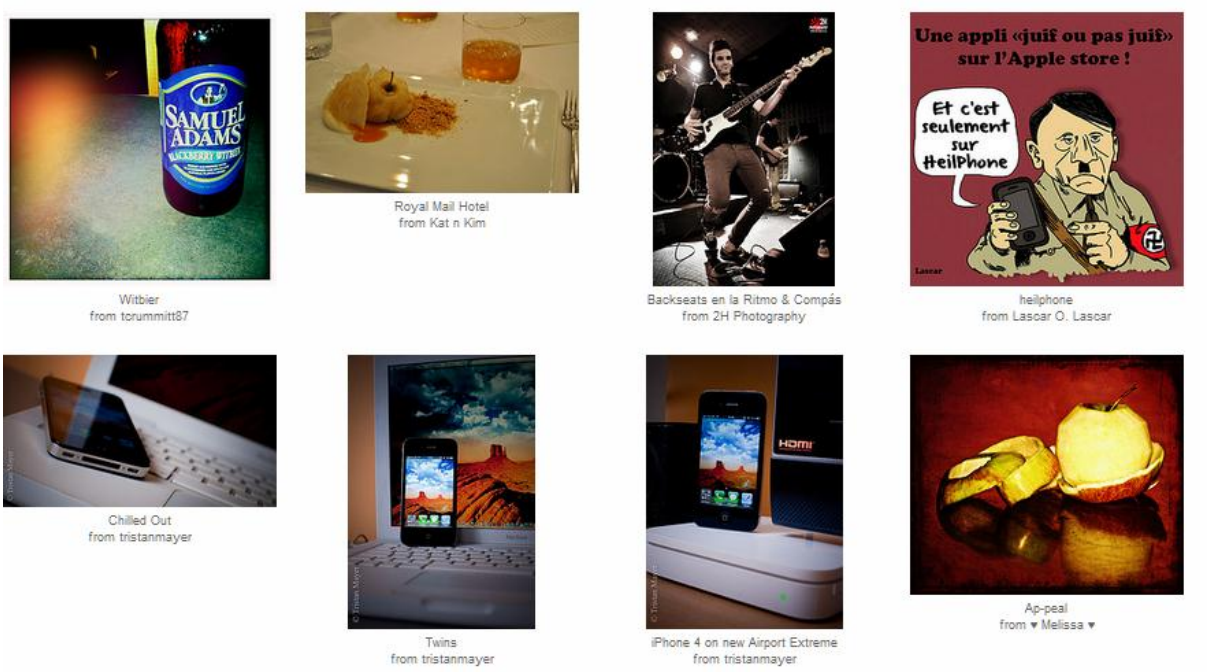

Figure 2-5: Tag ambiguity in a Flickr search for pictures tagged 'Apple'

Such ambiguity of tags is an existing issue in tagging systems where tags are a primary source of information retrieval (Golder & Huberman 2006). Therefore such systems are unable to differentiate the different contexts that are being pointed to by the same tag. Thus this polysemous nature of tags is potential problem and also affects the efficiency of information retrieval and classification mechanisms.

2.4 Tag Disambiguation

2.4.1 The problem

The issue of polysemy has been studied in numerous fields of applied sciences particularly in computer science where machine reads text without knowing the sense in which it has been used. Several disambiguation methods have been proposed in literature and each has its pros and cons. It is therefore significant to identify such technique that is efficient in using social network information while resolving synonymy or ambiguity of keywords.

2.4.2 Related Work

In computational linguistics the issue of ambiguity is addressed under the area called Word Sense Disambiguation (WSD) which is an open problem of natural language processing (NLP). WSD deals with the process of identifying suitable sense of a polysemous word occurring in a sentence. This technique helps improve the efficiency of applications concerning with machine translation, information retrieval and hypertext navigation, content and thematic analysis, grammatical analysis, speech and text processing (Ide & Jean 1997). Furthermore WSD has been classified to be n AI-complete problem, making it as challenging as solving central problem of artificial intelligence e.g. the Turing Test. Scientists believe that this known difficulty of WSD does not arise from a single

factor, number of reasons cause this such as the challenge of indicating word sense representation (including finite set and rule based senses) and coarse-grained lists of senses covering homonyms. One important factor about WSD is that it is considerably depends of external source of knowledge. The steps in any WSD algorithm roughly involve a set of words, a method which benefits from one or more sources of knowledge (e.g. machine-readable dictionaries), employed to identify and allocate suitable senses to words. The knowledge base is an essential part of WSD and it would not be possible to achieve the task of WSD without a suitable source of knowledge.

2.4.3 Approaches and methods

Several techniques have been developed to approach this problem. These can be divided into four main categories (Ide & Jean 1997) and (McCarthy 2009):

1. **Dictionary and knowledge-based methods**:
   Knowledge sources act as main component in this approach. These methods primarily rely on dictionaries, thesauri, and lexical knowledge bases, without using any corpus evidence.
2. **Supervised methods:**
   These methods employ training from sense-annotated corpora. A corpus provides a group of samples which allow the development of numerical language models, and thus these corpora are used with empirical methods.
3. **Semi-supervised or minimally-supervised methods:**
   This approach trains on a combination of knowledge base and small annotated corpus. The combination acts as a seed data to bootstrapping process or a word-aligned bilingual corpus.
4. **Unsupervised methods:**
   Unlike supervised methods, in which senses for a target word are selected from a closed list based on a dictionary or lexicon, unsupervised WSD tries to induce word senses directly from the training data. They almost avoid any source of external information and perform disambiguation from raw corpora. This technique is also referred to as word sense induction (WSI).

2.4.4 Unsupervised Word Sense Induction:

2.4.4.1 Advantages

Unsupervised WSI have a number of advantages over conventional methods. McCarthy (2009) summarises some of them as follows:

- No need of predefined knowledge sources and word contexts for disambiguation.
- Unsupervised WSI allow the corpora to be dynamic, as opposed to supervised methods.
- WSI systems can detect new senses from corpus data whereas systems using supervised methods are restricted to whatever sense distinctions are provided by the lexicographers.

2.4.4.2 Techniques

The challenge of WSD is not new and therefore number of techniques has been devised over the period of time which we will review in this section. Broadly we can identify three main techniques from literature namely:

- Context Clustering
- Word Clustering

- Co-occurrence Graphs

2.4.4.3 WSI using Context Clustering

In context clustering vector space model depicts a word. Each vector of a target word is clustered into groups, identifying a sense of a word. Schütze (1998) is his work proposes a similar technique which groups the occurrences of an ambiguous word into clusters based on the similarity between context and occurrence. Expectation Maximization (EM) algorithm does the clustering and cosine between the related vectors determines the context. Agglomerative clustering technique merges single cluster with similar pair of clusters (Purandare & Pedersen 2004). It continues merging with less similar pairs until a threshold is reached, authors report successful evaluation of results and claim their clustering algorithm to be effective than the former procedures.

2.4.4.4 WSI using Word Clustering

Word clustering involves finding words that are similar to the target word and using the clusters of words to convey a specific sense. One of initial studies on word clustering (Lin, 1998) identifies words w = (w1, . . . ,wk) similar to a target word w0. They calibrate the similarity between w0 and wi on the basis of information content of single features, obtained by syntactic dependencies, such as subject-verb, verb-object, adjective-noun, etc. With the increase in dependencies between shared words, the information content also increases. A subsequent approach by Lin and Pantel (2002) uses a different technique, called clustering by committee (CBC) algorithm. CBC automatically detects senses from text and avoids duplicating senses. The study shows that the evaluation of CBC outperforms many hybrid clustering techniques and manual evaluation of sample CBC outputs approves 88.1% of automatic evaluation.

2.4.4.5 WSI using Co-occurrence Graphs

This is slightly different approach to resolving tag ambiguity, it uses the concept of co-occurrence graphs G = (V,E) whose vertices V link to words and edges E attach the words having certain syntactic relevance. Veronis (2004) proposes a technique called HyperLex where co-occurrence graph is built comprising of nodes as words in text corpus, and a weighted edge between a pair of words is added to the graph in the event of co-occurrence in the same piece of text. The co-occurrence graph is then fed to an iterative algorithm where node having the highest degree is selected a hub. The process continues until the word conforming to a selected hub is below a fixed limit and this set of hubs represents the senses of a particular word. A similar approach (Agirre et al. 2006) based on PageRank reports the similarity between PageRank and HyperLex however PageRank have less parameters increasing its efficiency.

2.4.5 Existing Solutions

With advent of social Web, several computer scientists refer to this problem as tag sense disambiguation as (TSD) due to its obvious differences with traditional WSD. Albert Yeung investigates this area in great detail in his work (Yeung 2009) where numerous unsupervised methods are analysed and evaluated. One of the approach involves tag sense disambiguation through the analysis of the tripartite structure of folksonomies (Yeung et al. 2007), which is based on the GN algorithm. In this technique the graph comprising of resources linked with the target tag is divided into clusters (each representing one sense). The edge with highest betweenness is removed from the graph. The process continues until no more edges are left, making the division achieved

with highest value modularity. This division allows identification of clusters which act as the senses of the tag under consideration and the frequency of tags in each cluster helps in determining the signature of the matching tag sense.

Recent studies using knowledge based approaches also present an interesting perspective to TSD. Lee et al. (2009) use Wikipedia as a source of tag vocabulary and each occurrence of tag refer to a topic in Wikipedia. The algorithm involves identifying local and global neighbours with the help of co-occurrence relations, results in finding the best mapping from the occurrence to the Wikipedia topic by calculating relevance values between context and all topics. Analogous to this study, Garcia-Silva et al. (2009) used DBpedia[13] in place of Wikipedia as knowledge source. DBpedia is the structured and to some extent machine understandable form of Wikipedia. The study involves finding similarities between tag context and tag sense denoted by the bag-of-words model. Each bag corresponds to a topic entry in DBpedia for selecting the suitable mapping from tag occurrences to tag senses. Evaluation of results achieved from co-occurrence/clustering techniques show a better rate of disambiguation as compared to techniques using statistical models or temporal contexts. The process of evaluation lacks a 'Gold Standard', however the techniques employ manual classification of results. One other method to evaluate an algorithm is to test it one of the applications. For example information retrieval is one area that can hugely benefit with this disambiguation process.

2.5 Tag Disambiguation and Information Retrieval

Information retrieval is an area of computer science that is under constant research in order to make the user Web browsing experience an efficient and more personalised. As we have already established that user's behaviour on the Web is affected by online social interactions (Singla & Richardson 2008). This implies a need of a more smooth and customized Web browsing session for users. Many scholars have embarked on research projects concerning ambiguity free Web search. Kato et al. (2008) summarise the work done in improving Web search by using tags and urges that social tags can help improve the information retrieval, particularly from Flickr. Their method involves replacing the abstract keywords such as 'spring' with a set of concrete terms. Extraction and replacement is done by clustering the tags in accordance with the term co-occurrence of images. Experimental results show improvement in the recall ratio of Web image searches. Yeung at al. uses K-nearest classification on the returned clusters from the Web search engine (Yeung et al. 2008). They argue that instead of identifying different senses of the query terms, it is better to cluster the resources returned by the search engine. Each cluster corresponds to a context and fed to classification algorithm. The algorithm returns labelled classes in decreasing order of their rank. The k-nearest algorithm is effective as it handles redundancy of contexts when one or more contexts refer to the same. Yeung et al. (2008) employ this technique to extract semantics from a folksonomy (Delicious) to solve the tag ambiguity in Web search. They report to identify some unconventional meanings of polysemous words; however their work lack personalised search classification based on some criteria. Therefore it would be a novel idea to fill in this gap by inducing social network information of a user and give personalised results to every user.

[13] http://www.dbpedia.org

2.6 Summary

We have studied two of the most practised services of the Social Web; social networking sites and social tagging systems. Studies show the evidence of sociological traits such as influence, correlation, and also show strong consensus on extraversion in SNS. User behaviour is affected by these characteristics on individual and communal levels and the information present on the social network convey actual personality and inclination rather than self-idealization. Homophily is another important factor which governs the principle of information localization throughout the share community. Moreover, status and value homophily play important role in forming individual and collective opinions (Utz 2010). People express their interest in terms of free-form keywords which help in information organization, information retrieval, and contextualization. Facebook and Flickr represent prominent example of this social activity formally known as folksonomy, however polysemy is one of the issues that affects this social bookmarking activity. Ambiguities in keywords greatly degrade the performance of information retrieval systems.

The problem of keyword ambiguity is studied in many interrelated disciplines of computer science such as computational linguistics, artificial intelligence and information systems. Word sense disambiguation (WSD) broadly caters the area of synonymy or ambiguity. (Liu et al. 2010) and (Ireson 2010) summarise the work done in resolving this problem. Academics have employed supervised and unsupervised methods using statistical, knowledge based and clustered procedures. Although comparison between them is unjustified, the studies using unsupervised methods are more dynamic in identifying different contexts and also reduce human efforts. Therefore we can establish from previous work that unsupervised techniques perform better tag disambiguation in Web search sessions. Clusters identify the context and classifiers enable filtered results. We can also conclude that tag co-occurrence technique and k-nearest-neighbour algorithm show better results in terms of capturing social contexts as they handle the issue of context duplication and redundancy.

3 Methodology

The purpose of this study is to disambiguate polysemous words with the help of social network information. Therefore we have developed a comparative evaluation approach to test our hypothesis. The methodology is a combination of questionnaire and experimental comparison of state of the art algorithm. Initially we ask user what is the actual context when he/she is searching for a particular keyword, this gives us a baseline metric for evaluation purposes described in the later section. We then ask for user consent to gain access of the social network information of the user and his/her friend's. Storing this information, we then feed this data to our disambiguation framework. This concludes the experiment and the results are then evaluated manually against baseline metric to verify the hypotheses. The following Figure gives an overview of undertaken methodology.

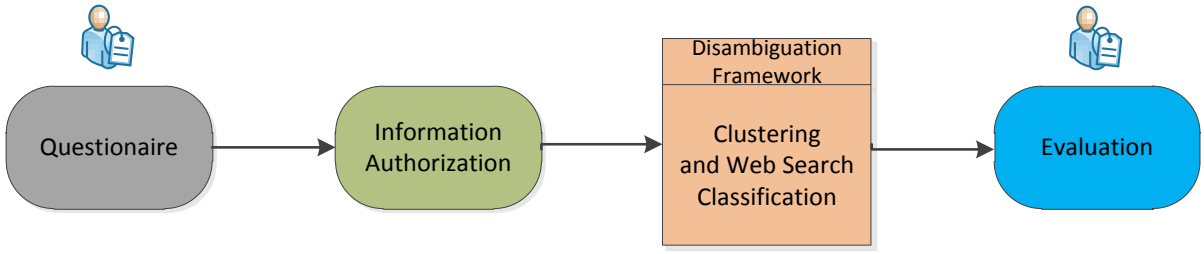

Figure 3-1: Overview of Methodology

3.1 ECS Ethics Committee Approval

As the study needed access to social network information of users, prior approval had to be taken from school ethics committee. An 'Expedited Limited Subject Data' form was filled with supporting documents such as project description, consent form and sample questionnaire. It was explicitly mentioned in the consent that no information that can identify users will be recorded and every bit of information will be stored anonymously on university computers.

Furthermore an email having all the mentioned documents was sent to the ECS Ethic Committee for formal approval along with signed application form by the investigator and supervisor. Application was granted approval on 18th August 2011 with reference number **ES/11/08/008.**

3.1.1 Participants

As we were approved to conduct experiments using limited subject data category, the participants had to be from University of Southampton, primarily from School of Electronics and Computer Science. We advertised this research on University's official Facebook Pages administered by student services. 20 students having Facebook profiles were selected to participate in this study. Participants ranged from Bachelors to MSc with mixed socio-demographic orientation as it was one the aspect of the research to study value and status homophily.

3.2 Questionnaire Analysis

Questionnaire was not conducted in a conventional way; it was added to the web application made to gather social network information. It mainly consisted of one question:

"If you were to search for images on Flickr or any other search engine, what would you likely be searching for if you typed the following?"

The user was asked to select one option out of 5 for each ambiguous tag, the radio buttons correspond to relative values accordingly. These values will be used in the evaluation section where we compute the efficiency of the algorithm with this baseline metric. Table 1 shows a key to selecting the answers in the questionnaire.

			Ambiguous Tag			
Meaning A	**O**	**O**	**O**	**O**	**O**	Meaning B
	Certainly	Maybe	Both	Maybe	Certainly	

Table 3-1 : Key to Selecting Radio Buttons

3.2.1 Selection of Ambiguous Tags

The process of selection is sig nificant as we are conducting the study on limited users with conforming interests. However the users belonged to diverse socio-demographic groups. The tags were selected from literature (Yeung 2009) keeping these aspects in mind. As the experiment include disambiguation of tags associated to Flickr photos, so ambiguous tags from Flickr were also used in the experiment. Sample tags with their possible meanings are shown in the following table. In this study we have taken the top 2 possible meanings of an ambiguous tag for valuation purpose.

1	**SF** Science Fiction **O O O O O** Sanfrancisco
2	**Bridge** Architectural Structure **O O O O O** Card Game
3	**Tube** Video Sharing **O O O O O** Mode of transportation
4	**Opera** Music **O O O O O** Internet Browser
5	**Language** Computer **O O O O O** Human

Table 3-2 : Sample Ambiguous Tags

3.3 Information Authorization

We have used Facebook as the source of social information of a user. The idea behind information access is to use it in disambiguation technique as it represents user's social characteristics and inclinations. Facebook profile has sections where people add information about their interests, activities, art and philosophy, and list of inspirational people etc. All this information is expressed in free form keywords and can be viewed by friends who can also add these keywords in their profile making it a collaborative effort. Therefore after filling the questionnaire, we ask the user to grant us access to his/her Facebook profile.

3.3.1 Data Collection

As the study is experiment based, data is of pivotal importance as the results and evaluation can completely change if erroneous data is fed to the algorithm. Therefore information that would best describe the sociological depiction is required in our case. Moreover, the information collected must be stored with referential integrity for efficient development and retrieval purpose. Data was collected from Facebook and stored anonymously in the database. The data consisted primarily of

user 'Likes' and friend's 'Likes'. The data collection was accomplished using Microsoft technologies with Facebook Open Graph API[14] and stored using SQL scripts in string data type (varchar). We explain detailed information extraction in the implementation chapter.

3.3.2 Facebook 'Likes'

Facebook 'Likes' have benefited the social Web is many ways, the 'Like' button is the most famous application that helps in page rank and credibility analysis. Moreover users can like almost anything on Facebook for example page, photo, comment, posts etc. However these 'Likes' are contributing more towards collective intelligence by enabling users to add diverse information in form of keywords (freely chosen of defined collaboratively) . This information comprises of varied perspectives such as likes, dislikes, hobbies, school of thought, professional and academic choices etc. The Graph API lets us view the list of 'Likes'[15] of a user, where every 'like' has a category which is defined by users and then it grows with the number of likes and recommendations from Facebook. This information is in raw form and has some issues to be handled like foreign language, special characters and alpha-numeric instances.

3.3.3 Data Cleaning

The experiment deals with disambiguating tags when queried to Flickr. This activity can suffer if we feed the raw data (Likes) to the algorithm therefore after gathering data from Facebook, it should be cleaned as it has noise in form of special characters, non-English language (uni-code). We have applied methods to remove special characters from the 'Likes'. Chinese, Arabic and other foreign language keywords were removed as they were beyond the scope of this project. Alpha Numeric occurrences were also not included in the experiment.

3.4 Disambiguation Framework

This is the vital part of the methodology where the actual experiment is conducted. After the data cleaning and questionnaire analysis, we perform the experiment one by one on each tag with respect to user's provided information. The main steps can be summarised as follows:

1. First we query the ambiguous tag to Flickr for the purpose of identifying clusters, each cluster corresponding to a separate context.
2. The clusters are translated into classifiers using the K-nearest-neighbour algorithm.
3. These classifiers are compared with the set of user's keywords and their similarity is valuated.
4. The classifiers then order themselves in descending order of similarity and each classifier has a set of tags attached to images.
5. Steps 1-4 are performed for all the ambiguous tags considered for the study.

Accordingly we have conducted the experiments in this manner on the information provided by 20 participants. Results are evaluated against the information gathered in the questionnaire and also through manual examination of images returned by Flickr.

[14] Facebook Graph API-https://developers.facebook.com/docs/reference/api
[15] Your Facebook 'Likes'- https://graph.facebook.com/me/likes?access_token={}

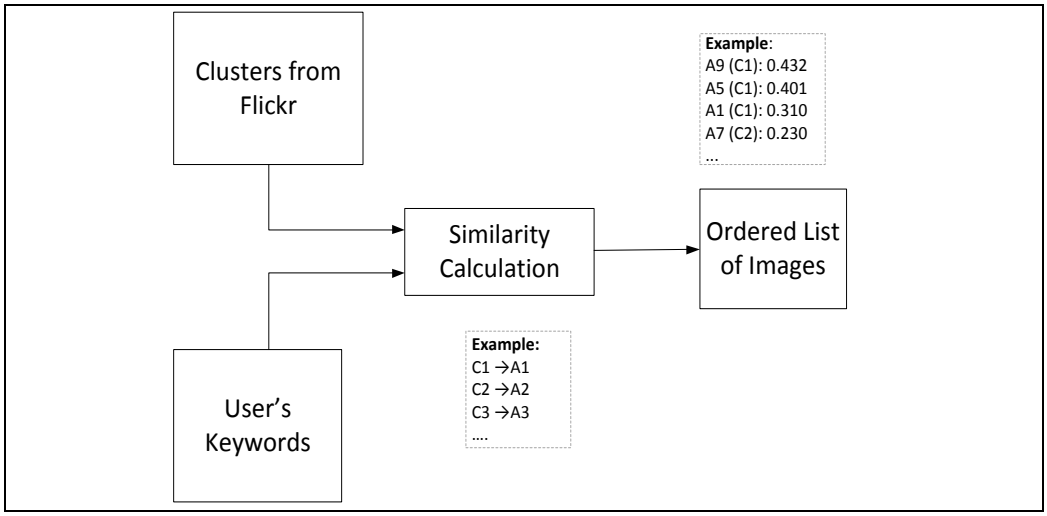

Figure 3-2 Overview of Image Categorization

3.5 Evaluation

The hypotheses will be validated in this final part of the methodology. We have used quantitative and qualitative analysis to verify our claims. The actual meanings of tags asked in the questionnaire are used as the ground truth for the evaluation of our approach. Moreover manual examination of images returned from Flickr gives us a probabilistic measure for calculating the efficiency of the algorithm. We perform manual examination by making three lists of images returned by Flickr about a particular tag. Each list contains 10 Flickr photos specific to the ambiguous tag.

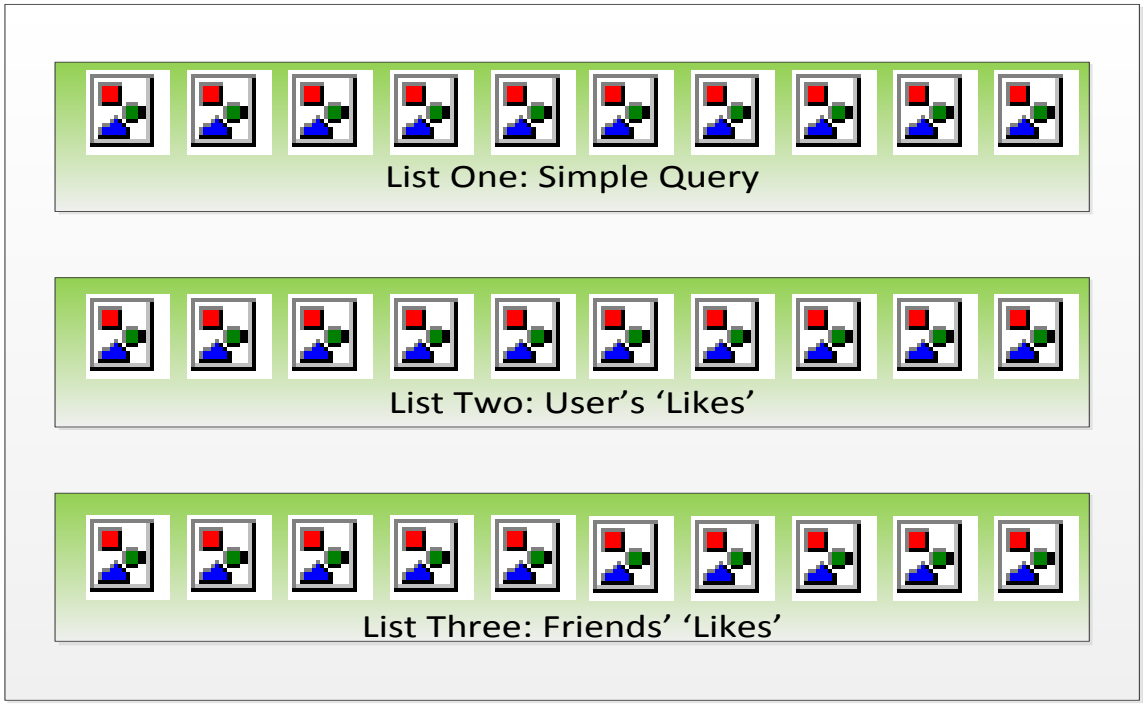

Figure 3-3: Evaluation Overview

3.5.1 List One: Simple Query

This list will include images returned from Flickr using the ambiguous tag without any filtering or manipulation. For example Figure 2-5 shows a list of images returned from Flickr with the tag 'Apple'. The images are queried and displayed in the order as they are on Flickr.

3.5.2 List Two: User's 'Likes'

Images returned from Flickr about the specific tag are fed to the Disambiguation algorithm where clustering and classification is performed and images are reordered according to user's social network information.

3.5.3 List Three: Friends' 'Likes'

This is the list where we test our hypothesis regarding homophily. We take the social network information of user's friends and disambiguate the images. The process is similar to list 2 however the data used is different and does not include user's 'Likes'.

3.5.4 Analysis

After populating the three lists we validate our hypothesis that the list 2 contains images related to what user has said in the questionnaire. Moreover our second hypothesis states that the list 3 should be better than list 2 with respect to images related to user's meaning. For this purpose we manually examine each photo and calculate the results in terms of photos matching to user input.

3.6 System Architecture

The architecture of the system majorly comprise of three components; information extraction and storage, disambiguation activity and evaluation. Figure 3.2 shows a high level architecture of the system mapping each component with its functionality. A website is used for the purpose of questionnaire and information extraction with the help of social network APIs. The stored information is cleaned and prepared for the experiment. The disambiguation activity comprises of acquiring the data from Database and performing clustering and classification on the images returned by Flickr. Results are analysed by manual evaluation of images and baseline metrics (List 1 and user input). The architecture diagram is considered vital for any project as it helps in the estimation of resources, implementation, time, and project management.

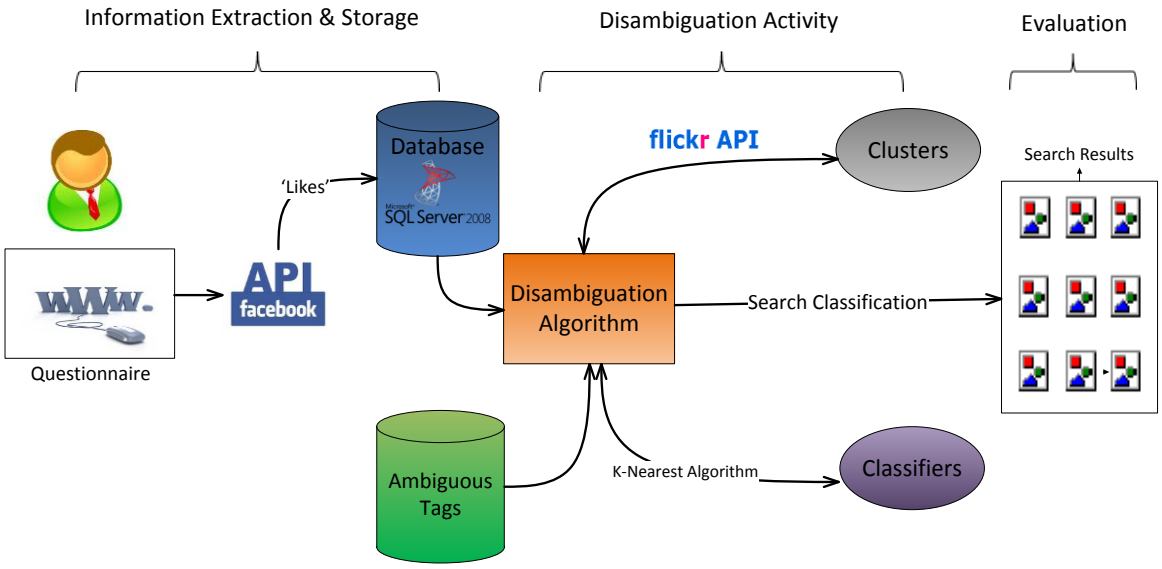

Figure 3-4 : Architecture Diagram

3.6.1 Estimation

After the agreement on the methodology with the supervisor, we estimated the time in terms of implementing the experiment from scratch to evaluation. The architecture diagram identifies the task and gives us a rough idea of how much resources are needed to complete the task.

	Task	Number of Days
1	Website Design and Questionnaire	2
2	Data Structures and Database Design	1
3	Facebook Api: Authorization and Information Storage	3
4	Flickr Api: Clustering Images	4
5	Classification Algorithm	5
6	Resolving Issues	2
	Total	17 (estimated)

Table 3-3 : Estimation w.r.t Implementation

3.6.2 Project Plan

While researching for the possible approaches to handle the problem, we found many techniques and methods which demanded in depth study. Activities like these and the system architecture

helped in making the revised project plan described in the following figure. To achieve these milestones, weekly meeting were arranged with the supervisor to discuss the progress and encountered issues. The social networks group in Learning Societies Lab also conducted meetings attended by most of the faculty staff supervising social networks group. In terms of project management, we used a tool called Mendeley which really helped in organizing all the research papers and documents. It also assisted in report writing as it has a plugin for MS word and can export citation in multiple formats. Traditional methods like maintaining a Logbook for different tasks was also used throughout the course of semester.

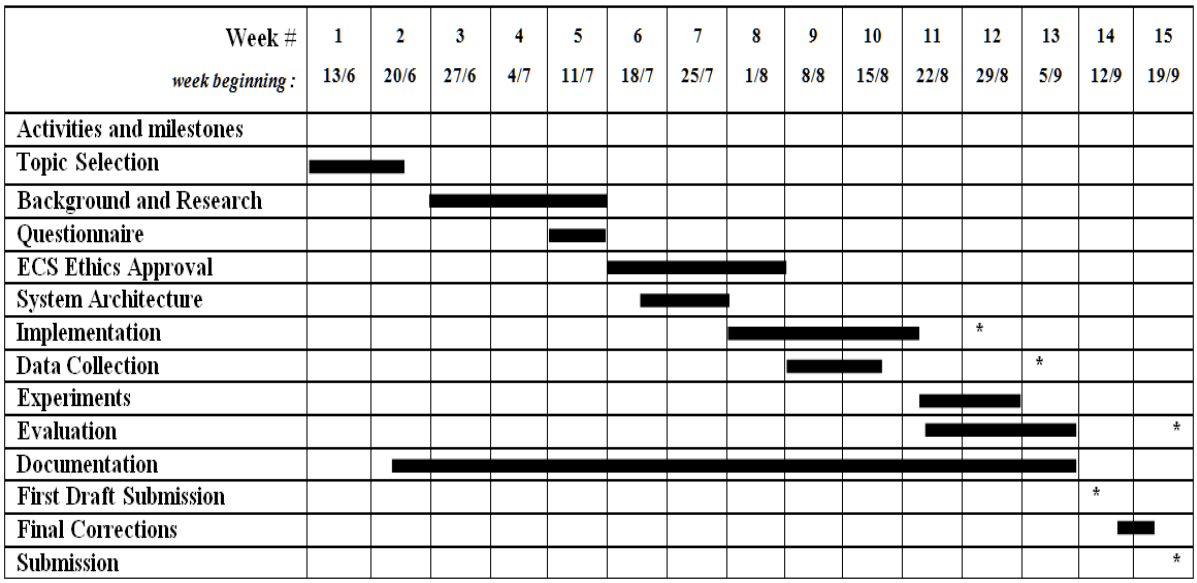

Figure 3-5 : Revised Gantt Chart

4 Implementation

4.1 ECS Virtual Machine

For the purpose of implementing the experiment, a virtual machine (VM) was requested from ECS Help Desk. The questionnaire was published on the website deployed on the VM. The VM was windows based equipped with Microsoft Visual Studio Ultimate 2010, SQL Server 2008 R2, Internet Information Services 7.5 and .Net framework 4.0. Special permission was taken for outside ECS access to the VM since the VM was only accessible within the ECS or on ECS VPN.

4.2 Technologies

4.2.1 Asp.Net 4.0 and C#

The website was designed using Microsoft technologies (Asp.net). The layout was reused from ECS website available to students of the school. The website was essentially developed for questionnaire fulfilment and information extraction from Facebook. Moreover, C# was used as the programming language for modelling all the logic and algorithms in the study. It should be noted that third party libraries and APIs were also employed for the experiment. Figure 4-1 and 4-2 show the implemented website having the consent information for the user with general instruction on how to fill out the questionnaire.

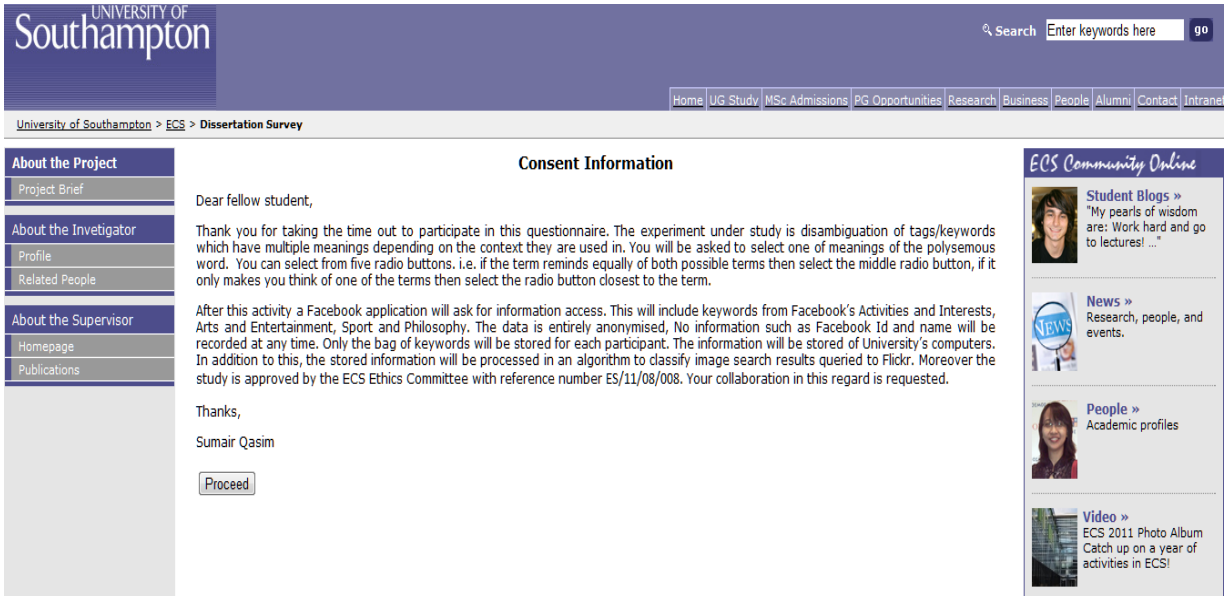

Figure 4-1 Website: layout and Consent Information

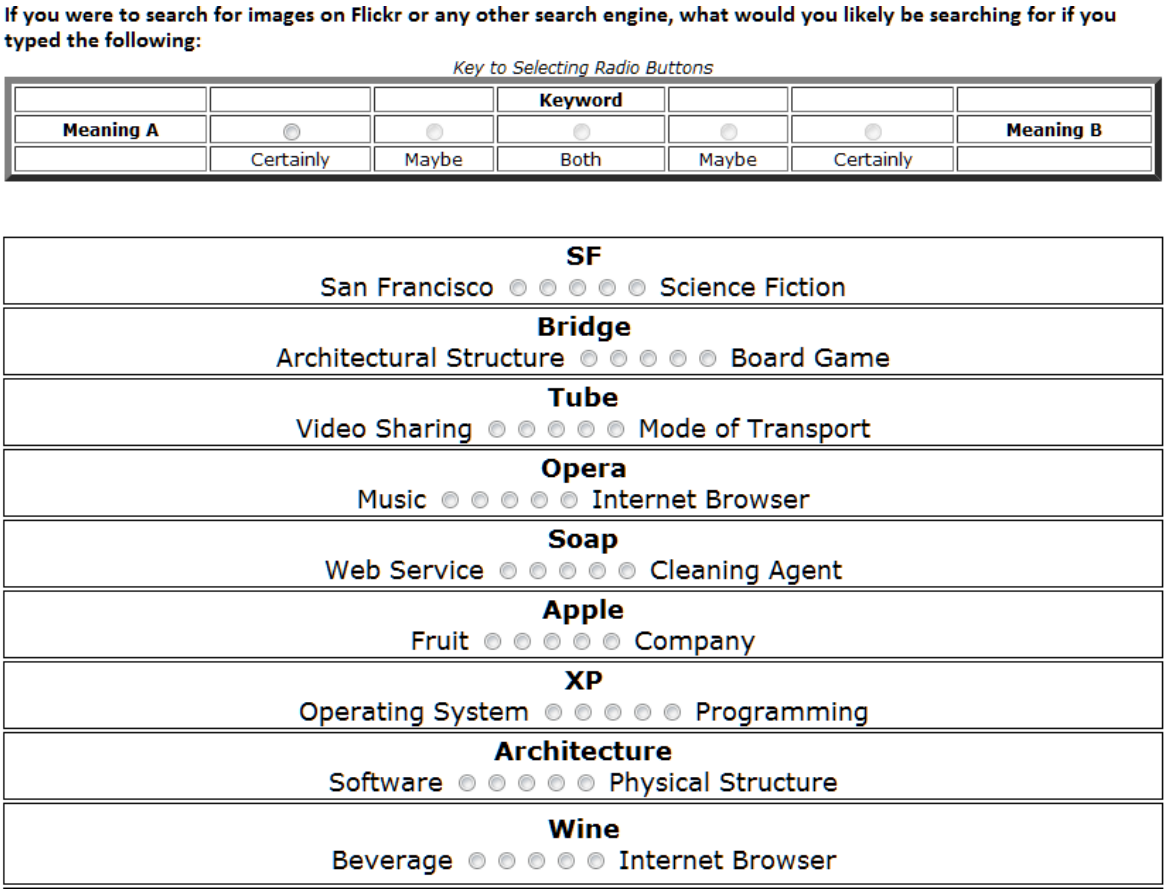

Figure 4-2 : Questionnaire with Key to selection

4.2.2 T-SQL

We have used MS SQL Server 2008 R2 for data storage using T-SQL programming with ADO.Net controls. Referential Integrity was modelled between tables called User (ID and Likes), Friend (ID and Likes) and Tag (ID and Meaning). The ID we are referring to is not Facebook profile ID, it was generated using a method.

4.2.3 Facebook API

After the questionnaire is filled, user is asked for information access to his/her Facebook profile. This was implemented our website using the Facebook Graph API. Extensive documentation is given on the Facebook developers'[16] website for implementing Facebook Authorisation, permissions and information retrieval. To this we have to first make an App[17] which generates an App key and Secret ID for authorization purpose. We also have to register the URL and Site domain while making the App as it redirects to the provided URL.

4.2.3.1 Facebook Authentication

We named the Facebook App as Dtags. Facebook allows authentication with OAuth 2.0, the application checks Logged In state of the user. If the user is already logged in, the login cookie is validated which is stored on the user's browser, authenticating the user. If the user is not logged in, they are prompted to enter their credentials as shown in the Figure 4-3.

[16] Facebook Developers-https://developers.facebook.com
[17] Facebook App https://developers.facebook.com/apps

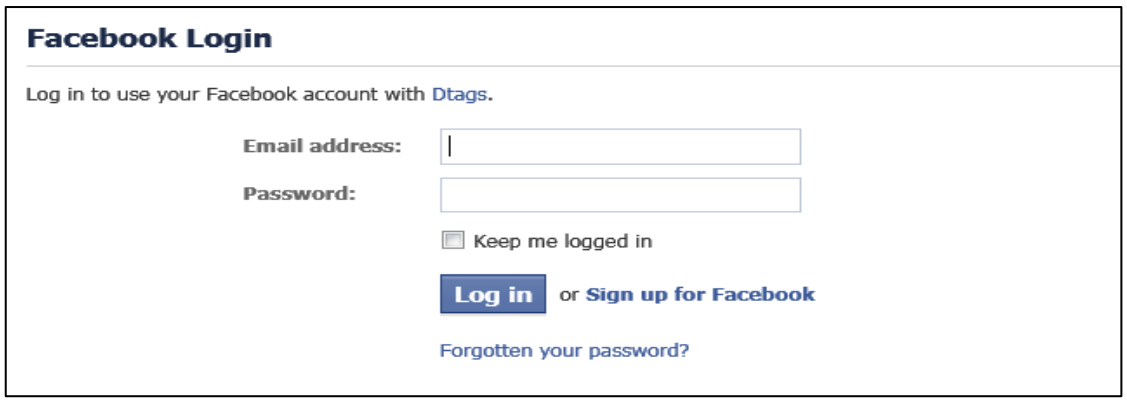

Figure 4-3 : Facebook Logn for Information Access

We implemented Facebook authentication with Asp.net and C# web programming. Facebook allow two types of user login; Client side and Server side, we implemented the Server side user login as our website was deployed on IIS webserver. The following URL is used for server side flow:

"https://www.facebook.com/dialog/oauth?client_id=YOUR_APP_ID&redirect_uri=YOUR_URL"

Client_id is the application ID generated by Facebook and redirect_uri is the address where the site redirects after success login. After the login is validated, the OAuth dialog asks for permission to access the data of user. According to Facebook documentation, there are 6 different kinds of permission that can be granted to Apps for data retrieval and sharing. We only needed user's information and list of his/her friends to access their 'likes' from the user's profile (provided friend has given access).

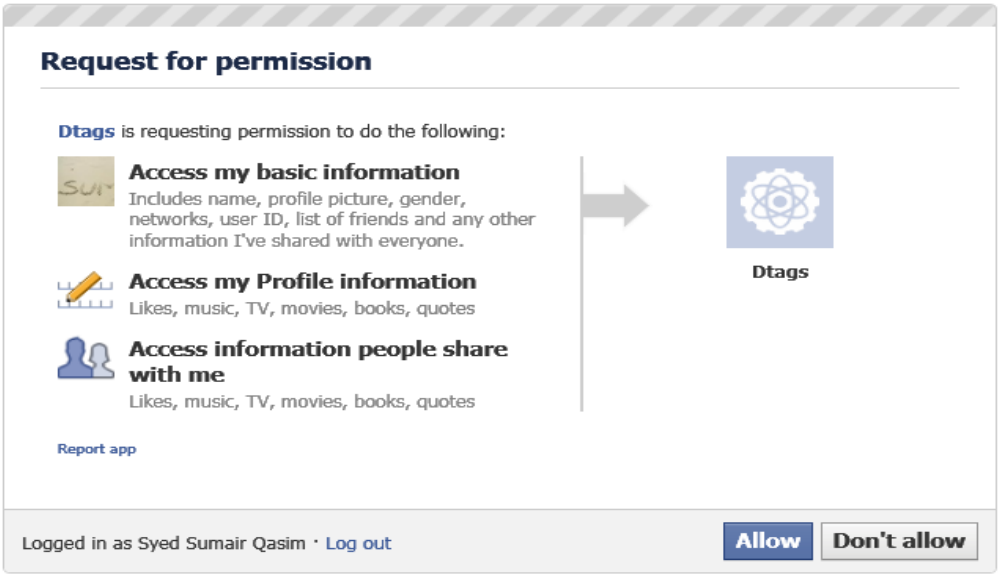

Figure 4-4 : OAuth Dialog asking for Permission

The following sequence diagram gives the overview of activities necessary for implementing the OAuth server side flow authentication. HTTP calls were used communication between Asp.net server (IIS) and Facebook through our application (Dtags). Once we have the token, we can make API calls to get user's information such as ID, list of friends, likes etc. (depending on permissions).

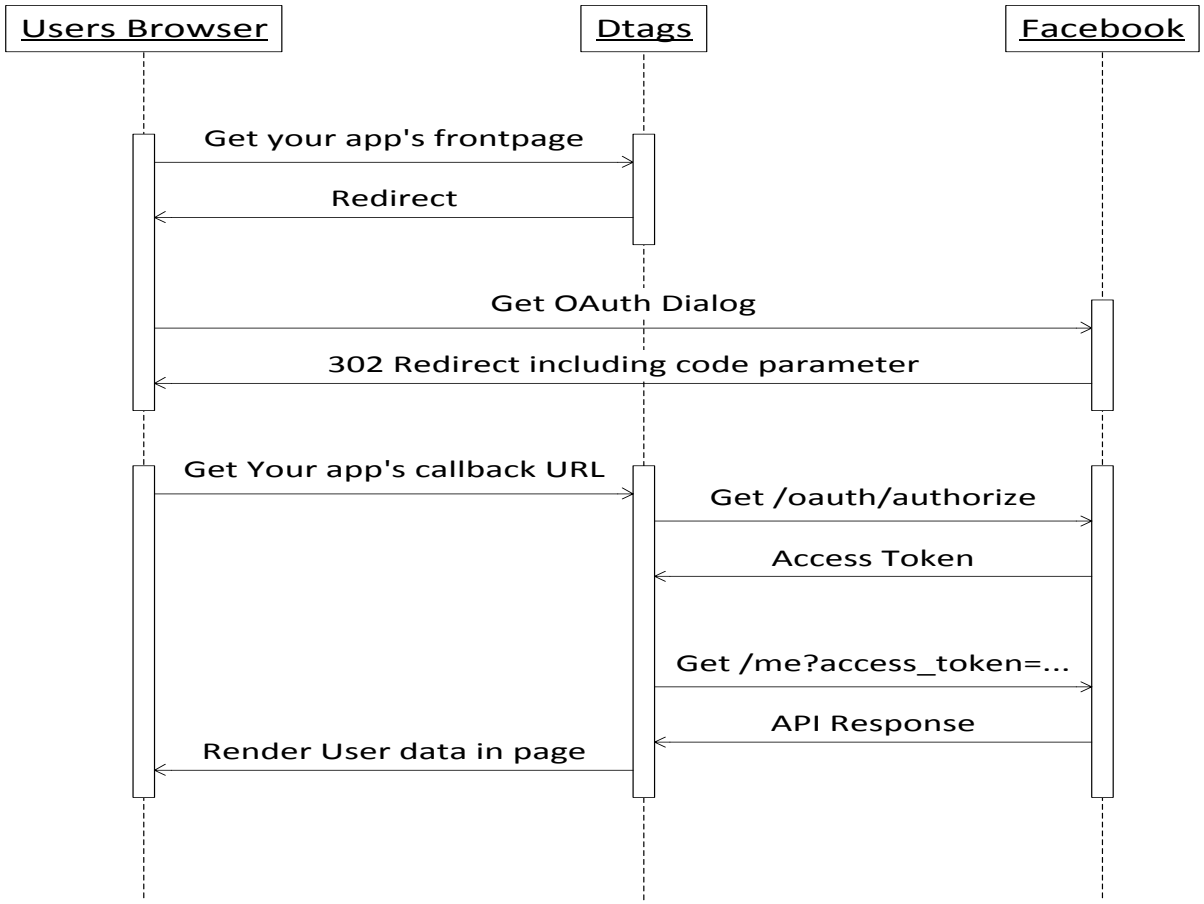

Figure 4-5 : Sequence Diagram (HTTP calls) for OAuth Authentication

4.2.3.2 Information Extraction

After the retrieval of token through Facebook authentication, we can now access the information by making calls to the Graph API. We used following API calls to get the 'Likes' of the user:

"https://graph.facebook.com/me/likes?access_token=" + oAuth.Token"

Here oAuth.Token is a local variable that stores the token acquired after the authentication. There is no call defined in the graph API to get the friends' 'Likes', therefore we first attained the list of friends by the following call:

"https://graph.facebook.com/me/friends?access_token=" + oAuth.Token"

Then we traversed each record and accessed their 'Likes' one by one. All data in Facebook is represented in JSON[18] format as shown in the following figure the Likes of user when called to Graph API.


```
{
   "data": [
      {
         "name": "Web Recruit LTD",
         "category": "Company",
         "id": "23574214300",
         "created_time": "2011-09-15T17:02:25+0000"
      },
      {
         "name": "Foundation Unversity Islamabad",
         "category": "University",
         "id": "278076255535829",
         "created_time": "2011-09-15T12:05:54+0000"
      },
      {
         "name": "Foundation University Islamabad (FUIEMS)",
         "category": "University",
         "id": "216745265002201",
         "created_time": "2011-09-14T18:01:35+0000"
      },
      {
         "name": "Amazon UK",
         "category": "Retail and consumer merchandise",
         "id": "136154403098698",
         "created_time": "2011-09-09T11:22:02+0000"
      },
```


Figure 4-6: Facebook 'Likes' JSON Format

4.2.3.3 JSON.Net

We used the JSON.Net (James 2007) library to traverse the JSON format while accessing the 'Likes', friend list and friends' 'Likes'. The Library has built in functions for parsing the JSON objects and converting them to string data type. We have encountered multiple issues while performing this activity and these are described in the next section.

4.2.3.4 Issues

We have come across many issues due to limitation of certain technologies, they are as under:

1. The JSON.Net format was raising illegal exception while storing in the database, it was due to the fact that some 'Likes' were in foreign language, uni-code and special characters. We handled this issue by applying Regular expression class, which removes such occurrences.
2. We also faced some HTTP exception while redirecting the OAuth dialog; this was due to incorrect declaration of User Agent. After some browsing of the Asp.net official forum this issue was also resolved.
3. 'Likes' of friends were acquired by using some nested loops, as we had to first get the list of friends, then find the 'Likes' of each user. This affected the efficiency of the website as we examined the trace of the website. For this issue, we divided the information extraction in few steps and used methods that were called after each step.

[18] JSON (JavaScript Object Notation) is a lightweight data-interchange format-http://www.json.org

4.2.4 Flickr API

To get the images from Flick we need to call the API for authentication and information access. Flickr also allows OAuth authentication like Facebook and the process of acquiring the token is quite similar to that of Facebook. First we have registered an App, which gives us Id and Secret Key. This helps in getting the token for making API calls. We make calls to Flickr API for getting images and clusters of those images. Flickr has provided documentation on its API that is for many programing languages.

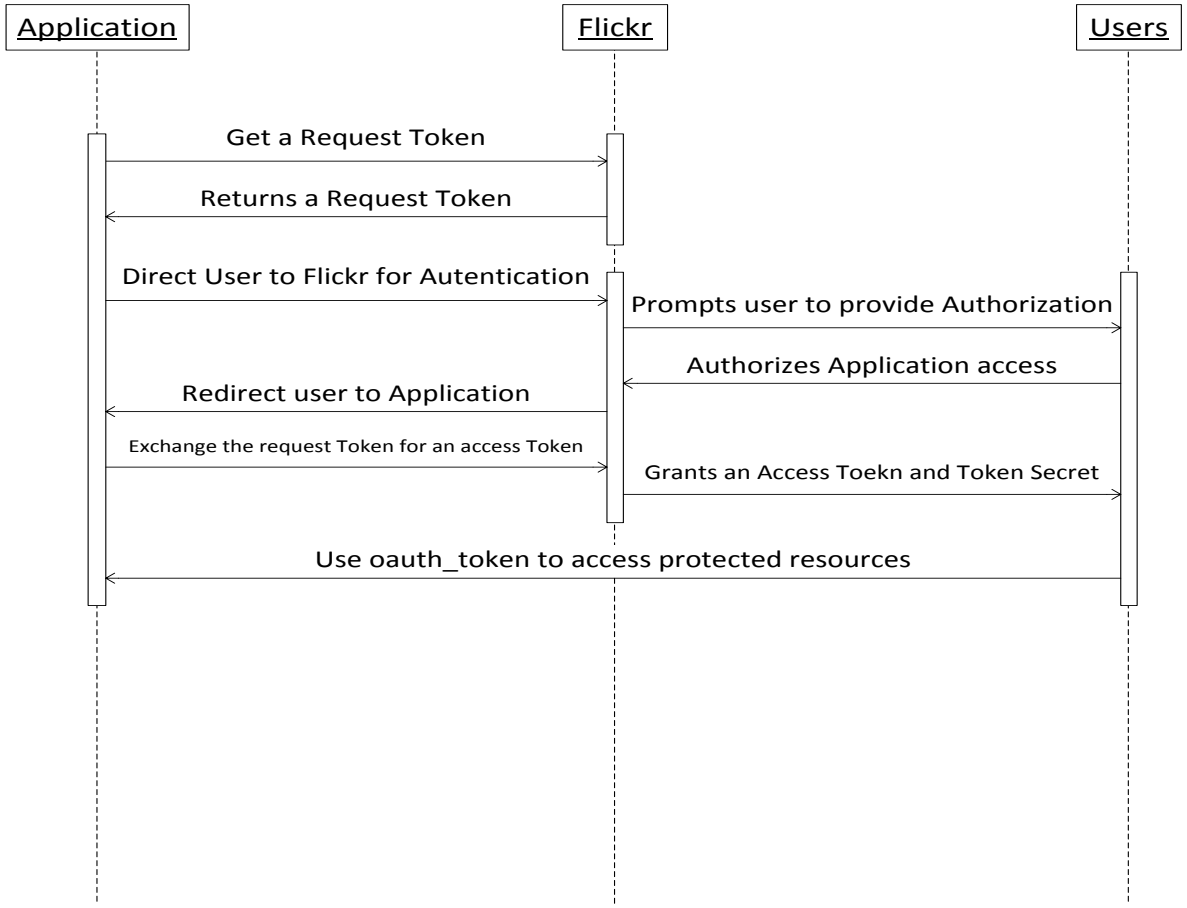

Figure 4-7 Flickr Authentication Overview

4.2.4.1 Flick.Net

Flickr.Net (Sun, 2010) is a third party library written in .Net framework for Asp.net web programming. We have used this library for authentication and token attainment, furthermore to

get images with their clusters. Flickr has its own algorithm for generating the clusters and each cluster represents separate context of the tag.

4.3 Clustering Images

The ambiguous tag is queried to Flickr to acquire the images for List One: Simple Query. We also use *Cluster class* available in the Flickr.Net API to find the clusters of the tag. Initially we obtain 100 photos from Flickr due to redundancy. Then we identify the clusters of the tag, Flickr has its own algorithm for clustering the images. Flickr.Net has *ClusterCollection* Class whose method is *TagsGetCluster(string tag),* used to call all the clusters. Figure 4-7 shows the list of retrieved clusters of all the 10 tags. Note that here each cluster here represents a context and contains tags within.

Tag	Clusters ID
Apple	nyc-newyork-manhattan fruit-red-green ipod-iphone-music mac-macbook-macintosh
Soap	bubbles-water-bath wash-car-carwash sapone-bolle-bolla handmade-natural-etsy
Architecture	city-art-bw sky-buildings-urban glass-blue-reflection building-london-nyc
XP	windows-mac-apple sky-green-grass xpro-lomo-film cross-process-processed
Sun	water-clouds-ocean blue-summer-beach sunset-silhouette-trees sky-light-nature
Wine	glass-red-bottle italy-italis-tuscany food-dinner-restaurant vineyard-grapes-california
Tube	london-england-uk underground-metro-subway macro-extension wave-surf
Opera	vienna-wien-austria sydney-house-australia paris-france-garnier oslo-norway-norge
SF	sanfrancisco-california-bridge
Bridge	night-city-sky newyork-brooklyn-manhattan water-boat-green river-london-italy

Figure 4-8 : Clusters from Flickr

4.4 Building Classifiers

We have used the approach proposed by Yeung (2009) where K-nearest-neighbour algorithm (KNN) is used to build classifier, although there is a significant difference in terms of data and folksonomy under study. To apply the KNN algorithm, we need to build classifiers who are less redundant. As it is obvious from Figure 4-8 that there are some clusters which share a degree of similarity between them. Consider the clusters returned by Flickr as:

$$F_T = \{F_{t,1}, F_{t,2}, \ldots\ldots, F_{t,m}\}$$

Where each cluster has a set of tags $T_{t,i}$ that makes the cluster (F_T). To handle the issue of redundancy we introduce a method of *overlap* that merges the two clusters if their tags ($T_{t,i}$ *and* $T_{t,j}$) are similar. We model the function as follows:

$$Overlap\ (T_{t,i}, T_{t,j}) = \frac{|T_{t,i} \cap T_{t,j}|}{|T_{t,i} \cup T_{t,j}|}$$

This method gives a ratio which we compare to threshold (α) and merge if the overlap function ≥ α. The merged clusters after undergoing this function are represented as:

$$C_T = \{C_{t,1}, C_{t,2}, \ldots\ldots, C_{t,n}\}$$

These clusters are referred to as the K-nearest classifiers returned when queried to Flickr.

4.4.1 K-Nearest-Neighbour Algorithm

The KNN can be summarised in the following algorithm. We input the clusters returned by Flickr pertaining a specific tag. We then extract tags of each cluster and compute the overlap and merge if is greater than or equal to a threshold value (α). New classes are formed after merging of similar clusters and these are called KNN classifiers. Following table shows an instance where the classes have been reduced to three using the KNN method. The algorithm adapted from Yeung (2009) can be summarised in Algorithm 1

Tag	Classes
Apple	nyc-newyork-manhattan fruit-red-green ipod-iphone-music-mac-macbook-macintosh

Table 4-1 : Classes Identified for Images Tagged with 'Apple' where α= 0.2

Input: Clusters F_T from Flickr for particular tag
Output: A set **C** of classes with a set of labels **T**

1. F ⇐ Flicker Clustering(tag)
2. T ⇐ {}
   {Extract frequent tags}
3. for F_i ∈ F do
4. T_i ⇐ *ExtractTags*(F_i)
5. **T** ⇐ **T** ∪ {T_i}
6. **end for**
   {Merge similar clusters}
7. merged ⇐ 1
8. **while** merged = 1 **do**
9. merged ⇐ 0
10. **for** T_i,T_j ∈ **T** and i ≠ j **do**
11. **if** overlap(T_i,T_j) ≥ α **then**
12. C_{new} ⇐ F_i ∪ F_j
13. **C** ⇐ **C** − {C_i,C_j}
14. C ⇐ C ∪ {C_{new}}
15. T_{new} ⇐ *ExtractTags*(C_{new})
16. **T** ⇐ **T** − {T_i,T_j}
17. **T** ⇐ **T** ∪ {T_{new}}
18. merged⇐1
19. **end if**
20. **end for**
21. **end while**
22. **return C,T**

Algorithm 1 Building K-Nearest-Neighbour from Flickr (Adapted from (Yeung 2009))

4.5 Web Search Classification

Now that we have obtained the set of classifiers in the previous section, the images can be classified according to user's social annotations. The k-nearest-neighbour classifiers represented as C_T are used compared with annotations ('Likes') of the user and their degree of similarity is calculated. The classifiers will be ordered in descending order of their degrees of similarity. Let the social annotations be represented as $A_{t,j}$ and tags(labels) related to a classifier as $K_{t,i}$. The similarity between them is calculated by the Dice Co-efficient and modelled with the method called *Sim*($X_{t,j}$,$Y_{t,i}$)*:*

$$Sim\ (A_{t,j},K_{t,i}) = \frac{2 \times \left|K_{t,j} \cap\ A_{t,i}\right|}{\left|K_{t,j}\right| + \left|\ A_{t,i}\right|}$$

In this manner, each set of tags of a classifier is compared with user social annotations ($A_{t,j}$). The similarity function returns the degree of similarity which helps in ordering the classifiers. Subsequently the images belonging to the classifier with the highest similarity are displayed in descending order. The flow chart of Flickr search classification is shown in Figure 4-9.

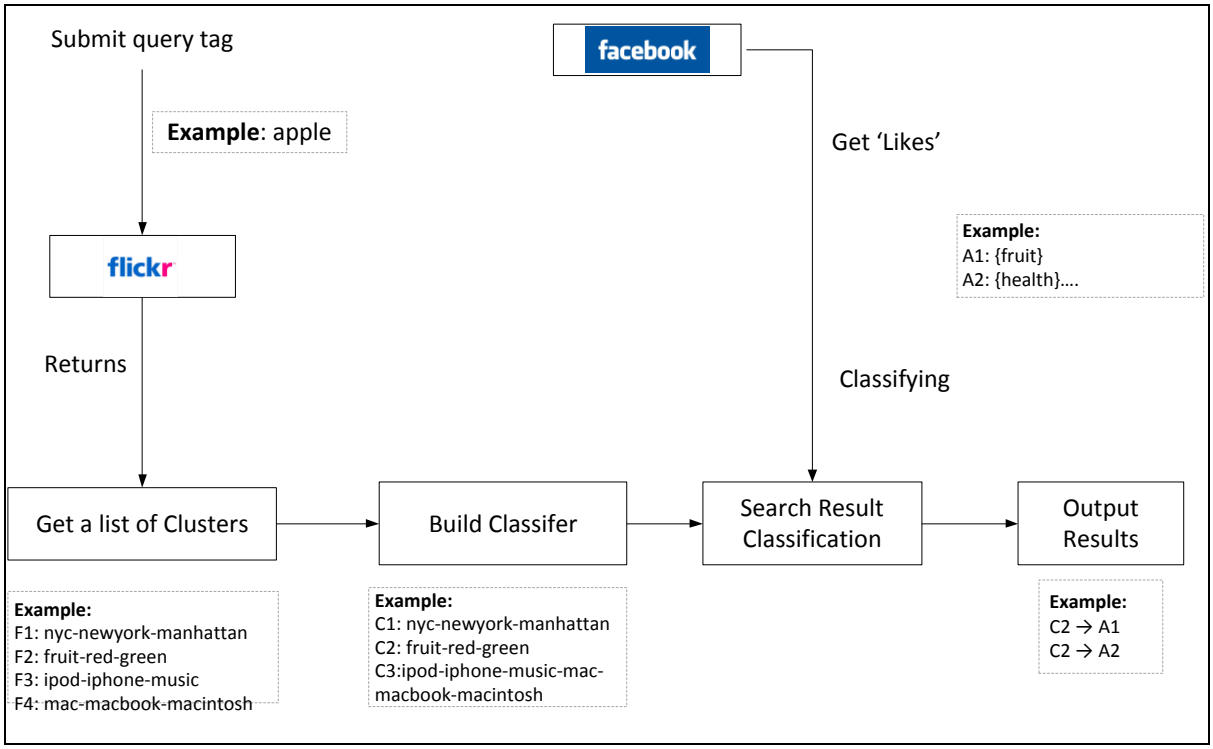

Figure 4-9 : Flow Chart of Image Classification

4.5.1 Issues

While classifying the images we came across issues where user has no 'Likes' related to the context under study. Similarly, the Friends' keywords can also exhibit the same nature. We handled this case by introducing a threshold (β). Therefore if the classifiers (KN neighbours) have the similarity value less than β, then they are considered as misclassified. One other issue was of efficiency of the approach in terms of time, as they were user profiles having considerably enormous amount of 'Likes' making the similarity method consume more resources in calculating degree of similarity. We have not approached this issue as it was not in the scope; however we did clean the 'Likes' from erroneous occurrences.

5 Evaluation

5.1 Experimental Setup

Our experiment includes generating the classifiers on the 10 tags that we have selected for this study. We call the clusters from Flickr related to each tag and apply the KNN algorithm to build classifiers. The value of overlap threshold (α) was taken as 0.3 meaning if more than half of the clusters are overlapping they will merge. These classifiers then compared to the set of 'Likes' of each user and are arranged in descending order of similarity. Images comprising these clusters are presented to user as returned search results. We had 20 participants mostly from ECS for this study. We conducted our experiment 20 times for each of the 10 tags to generate the second and third list, list one was generated by simply querying the Flickr Api with the tag.

5.2 Manual Examination

As disambiguation techniques lack 'Gold Standards' for evaluation, we made user interpretation as the ground truth while evaluating the returned images from Flickr. In the table 5-1 results of manual examination is shown for images tagged with 'apple' where context refers to the interpretation of the tag by the user, which we recorded in the questionnaire. It is vital to note that the Flickr is a dynamic database and is subject to frequent additions of images tagged with different keywords. Therefore we conducted the experiment on span of 5 days (22nd - 26th August 2011) and stored the results in html files. Moreover, the value of k was taken to be 4 as no more than 4 clusters were identified from Flickr. Misclassification threshold (β) was taken as 0.20.

User	Context	List 1: Simple Query	List 2: User's 'Likes'	List 3 Friends' 'Likes'
1.	Fruit	2/10 = 0.4	3/10 = 0.5	3/10 = 0.5
2.	Company	0.4	0.4	0.5
3.	Company	0.4	0.6	0.4
4.	Company	0.5	0.7	0.5
5.	Company	0.4	0.5	0.6
6.	Fruit	0.3	0.5	0.5
7.	Company	0.3	0.5	0.7
8.	Fruit	0.2	0.4	0.8
9.	Company	0.5	0.6	0.7
10.	Fruit	0.3	0.5	0.6
Average		0.35	0.5	0.6

Table 5-1: Manual Examination of Images tagged as 'Apple'

Table 5-1 shows the images examined of 10 random participants, their context and number of images correctly identified out of 10 returned images. This activity was repeated for the other 9 tags and results were tabulated manually (see Appendix). Moreover, figure 5-1 to 5-3 shows the outputs of the three lists; Figure 5-2 and 5-3 displaying the images returned using the algorithm where apple was used as the ambiguous tag.

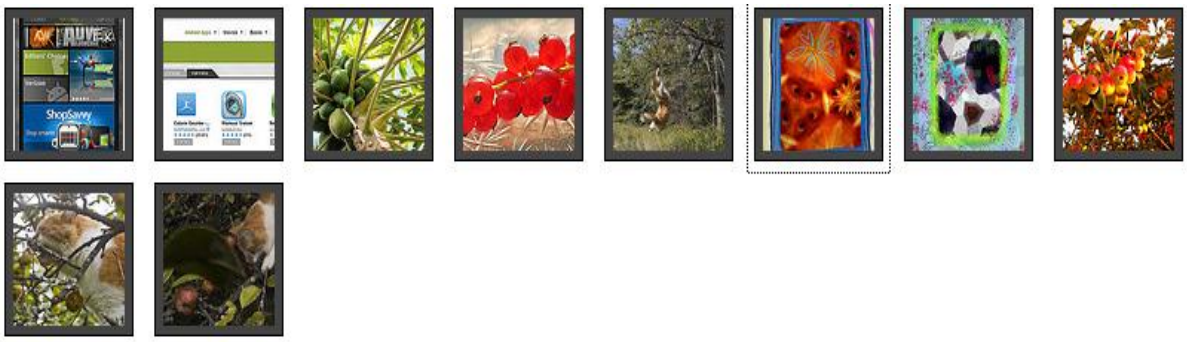

Figure 5-1 User 1: Output for List 1; Images tagged 'Apple'

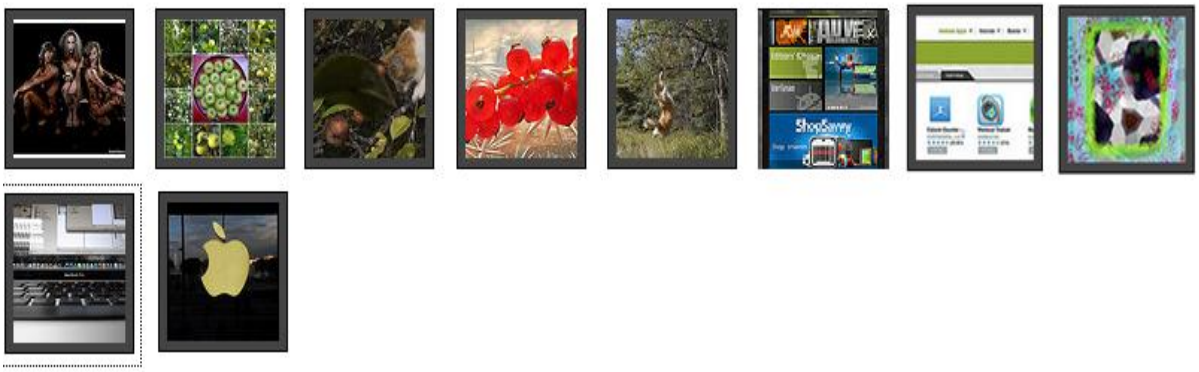

Figure 5-2 User 1: Output for List 2; Images tagged 'Apple'

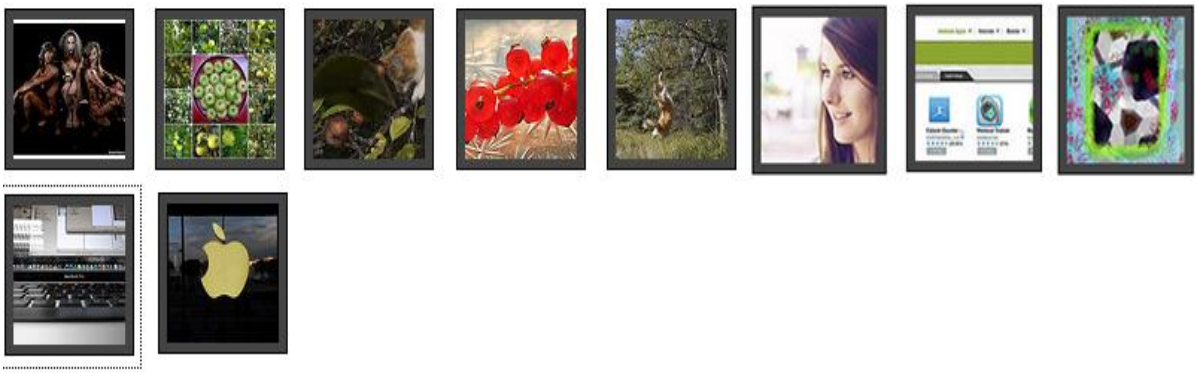

Figure 5-3 User 1: Output for List 3; Images tagged 'Apple'

After the conducting the experiment with the first tag, we present the preliminary analysis of results in the following graph. It is clear that accuracy of list two is better than list one (on average). The List three test the evidence of homophily and reports acceptable results. We call the ratio of relevant images and retrieved images as accuracy.

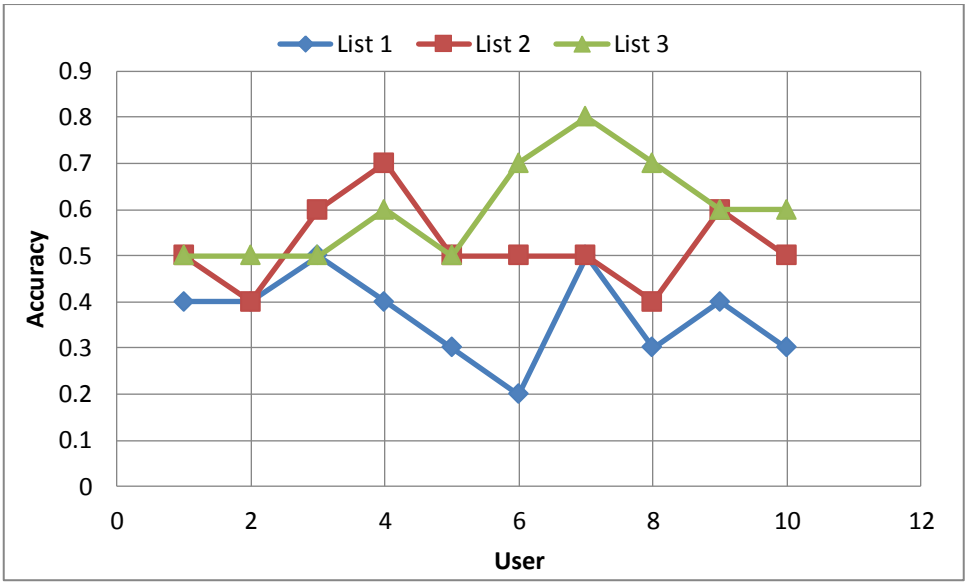

Figure 5-4 Preliminary Results of Manual Examination (tag=apple)

5.3 Results Interpretation

After formulating tables such as Table 5-1 for each of the 10 tag, we compute the average values of accuracy of all the tags. The average is calculated out of 20 users. Results can be seen in Figure 5-5, the List 1 (Facebook 'Likes') and List 2 (Friends' 'Likes' i.e. homophily).

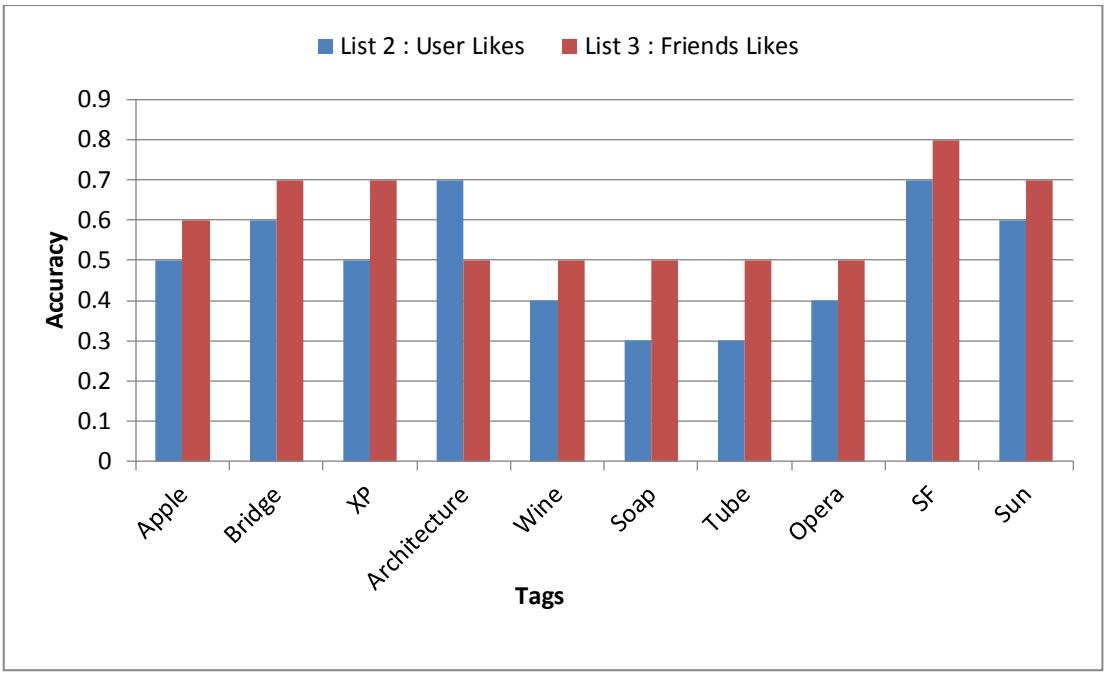

Figure 5-5 Accuracy of All tags after Manual Examination

After conducting the experiment on all the tags and manually examining the images to identify their context and accuracy, we observe satisfactory results conforming to our hypotheses. Facebook 'Likes' are able to disambiguate the polysemous tag, also establishing that there is a relation between social networks and online behaviour. According to the Figure 5-5 'Architecture' and 'SF'

are tags with highest accuracy (70%), meaning that our approach successfully returns 7 out of 10 images matching user's context. Moreover, the accuracy of list 3 is more than list 2 in most cases, validating the evidence of homophily its effectiveness in resolving uncertainty found in tag contextualisation. We analyse the performance measure in the following sections.

5.3.1 Quantitative Analysis

5.3.1.1 Precision and Recall

We use two performance metrics in order to quantitatively evaluate our approach. Precision measures the extent to which the images correctly classified. It is calculated by dividing the number of correctly classified images by the total number of images retrieved.

$$precision = \frac{|\{Correctly\ Classified\ Images\}|}{|\{Images\ Retreived\}|}$$

When precision is equal to 1, then no images are classified. Recall is a metric that computes the fraction of images that fall into one of the contexts (classifiers), it is calculated by the following formula:

$$recall = \frac{|\{Classifiable\ Images\ \}|}{|\{Images\ Retreived\}|}$$

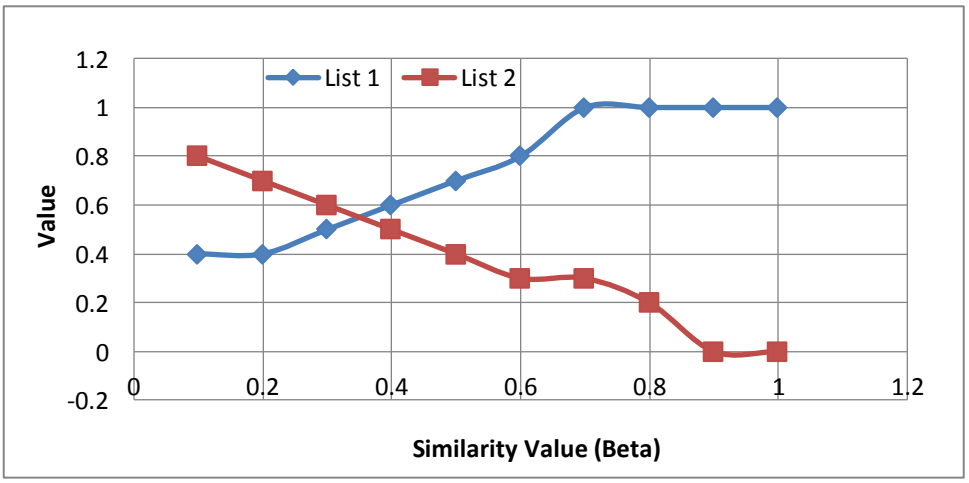

Figure 5-6 : Precision and Recall of Images Tagged Apple at Different Values of Beta (only for list 2)

If we take a closer look at the graph in figure 5-6, precision is affected positively with the increment of similarity threshold till a certain point, and then it becomes constant at 1.0 meaning no images are classifiable. This is due to the comparison of degree of similarity with its threshold, if no classifier is considered for viable enough then images will not be ordered (precision =1). However, we also observed that recall decreases with the increase in similarity threshold. This is because, when the threshold increases the number of classifiers decrease resulting in no classifiable documents.

Similarity threshold (β) played an important role while ordering the classifiers, we run our experiment using different values of similarity threshold (β) on the first case (apple) see Figure 5-6, and it was set to 0.25 for meaningful interpretation of precision and recall. With the help of formulas

defined above, we calculate the precision and recall of list 1 and 2; the values are displayed in the Figure 5-7 and 5-8 respectively.

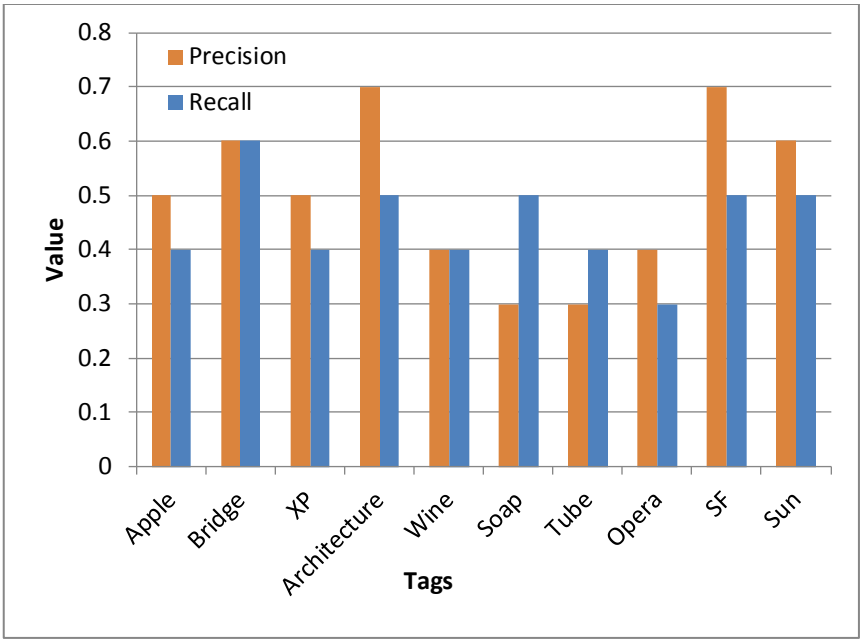

Figure 5-7 Precision and Recall of List 2 (User's Likes)

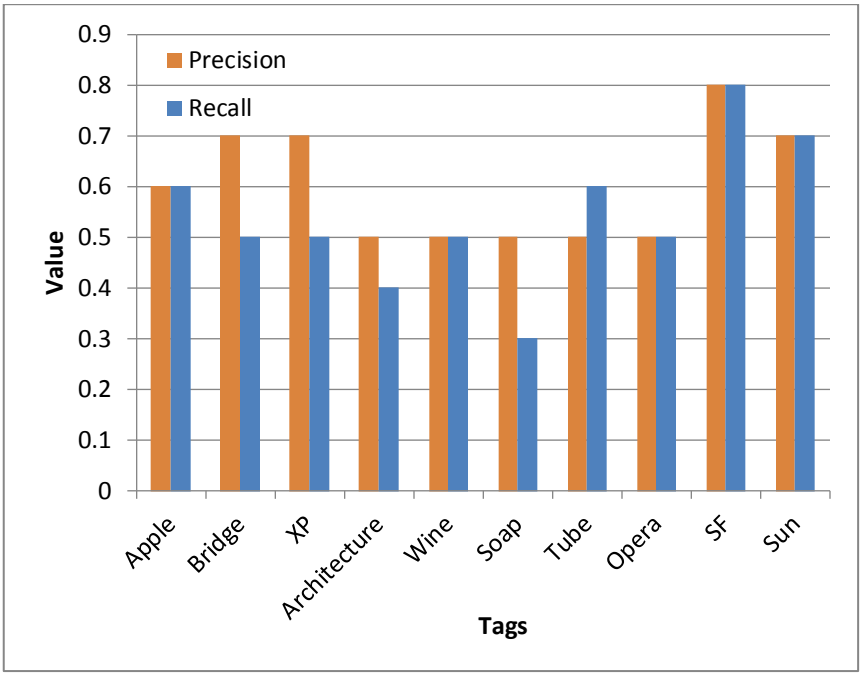

Figure 5-8 Precision and Recall of List 3 (Friends' Likes)

The values of accuracy gathered after the manual examination we performed in the previous section are actually equal to the precision values of both the lists. 'SF' is one common occurrence in both the lists that has been reported to have highest precision. Each value (precision and recall) in the above graphs is calculated by taking averages of experiments performed on the data of 20 users. The

number k-nearest-neighbour were taken as 4 and the similarity threshold was taken as 0.25 as observed from figure 5-6. The values acquired in figure 5-7 and 5-8 are the result of 400 manual examinations of images, 200 examinations for each list (see appendix).

5.3.2 Qualitative Analysis

Tag disambiguation is such a problem where extensive research has been conducted, yet there is no defined method of evaluation of results. Therefore we developed our own method of evaluation; the user context and manual examination of results. The two lists were examined separately reducing the effect of partiality. To start with, we examined the classifiers that are generated from Flickr clusters. These clusters represent the different contexts to which the images belong to. It was interesting to find that 'SF' has only one cluster in Flickr. Moreover, some of queried tags did not belong any of the contexts asked in the questionnaire, this effected the manual examination of the images, as we excluded these occurrences from manual examination, though they were a part of the classifier ordering activity and reported conformance with the user defined context. This was due to the presence of relevant keywords in the user's data We also found some level of similarity in the user's data, the reason being the fact that these 'Likes' are shared on 'News Feed' of friends whenever a person likes an object. Due to this activity on Facebook, the 'likes' are widely propagated within the friend network also localizing them throughout the homophilous network. Socio-demographic traits have also indirectly contributed to this study as the participants belong to one school of thought; however they were from quite different in their relative backgrounds. Thus the whole activity of disambiguation depends on numerous diverse aspects, addressing those can favourably affect the outcomes.

5.4 Discussion

The study of tag disambiguation based on social network annotations has revealed interesting facts in compliance with what we speculated. Investigation of results shows that our proposed strategy yields acceptable results. Both precision and recall at high values with a similarity threshold of 0.25 gives the accuracy of about 65 % (on average). The low similarity threshold establishes that the images are categorized even when few important keywords are present in the user's data. Moreover, it also describes that context can be identified with a limited but relevant amount of keywords.

In-depth analysis of result reveals the performance of our approach with respect to different tags. With 4 classifier and similarity threshold of 0.25 our proposed method gives the precision ranging from 30 % (Tube, Opera) to 70 % (architecture, sf) for user's keywords and 50%(wine, soap, tube) to 80%(sf) for friend's Keywords. We also investigate into cases that have lower precision and it is found that the keywords present in the user's profile are conflicting to their provided context. Moreover these are the cases where the keywords are more diverse instead of being relevant to the classifiers, hence disturbing the values of similarity function which orders the classifiers. Similar issue was discovered with images in Flickr, where some common tags in clusters disrupted the classifier ordering scheme.

5.4.1 Factors Affecting Results

Qualitative and quantitative analysis revealed some factors that if addressed can increase the efficiency of the approach.

1. Socio-demographic factors play an imperative role in studies where contextualisation is the problem at hand. As the participants in this study belong to a certain school of thought and share some common demographic traits, certain amount of partiality in results is reported.
2. The keywords in user's Facebook profile are wide-ranging in certain cases where precision is reported low. Moreover these keywords also point to more than one context resulting in incorrect ordering of the classifiers.
3. Flickr clusters do not contain the context which is indicated by the user. Moreover, some clusters have common tags.
4. Rare cases of much less keywords and few number of friends have been observed in user's profile.

5.4.2 Comparison with Related Work

We have reviewed the literature in the chapter 2 in terms of related research and application concerning tag disambiguation. As far as approach is concerned one of the recent works that is closely related to our study is the work by Yeung (2009). The scholar has extensively studied the issue of tag ambiguity in folksonomies such as Delicious. The approach includes building classifiers from folksonomies and then comparing the tagged documents returned by Google. In a way our strategy handles the issue the other way round Yeung's idea. In our case, we build classifiers on the clusters from Flickr when queried with the ambiguous tag and compare them with user's social network annotations. However the efficiency of our approach is reported to be less than the mentioned work, this is due to some factors identified in the previous section. One major reason that is worth mentioning is that our method orders the classifiers rather than the documents as in the other study, moreover we also employed the concept of homophily while disambiguating tags which shows promising results. Therefore we can imply that by handling some issues the performance of our approach can be increased significantly.

6 Conclusion and Future Work

In this study we have researched two of the most widely used services of the social Web, which are social networking sites and social tagging. We employ the characteristics of the former to address the weakness of the latter. Our research objectives included evaluating the effectiveness of social network information in disambiguating tags and we hypothesized that the information present in profile of a user reflects his/her interests and therefore can be employed to disambiguate tags. Moreover these interests when shared in a community produce a localizing effect called homophily and this phenomenon can also be exploited in contextualizing the ambiguous tags. To study the problem in detail we reviewed the relevant literature from diverse fields of applied and social sciences. For instance, we studied the sociological aspects of social networking sites and effect of these on user online activity. We also explored state of the art techniques and methods that are used to address the issue of ambiguity.

Studies show the evidence of sociological traits such as influence, correlation, and also show strong consensus on extraversion in SNS. User behaviour is affected by these characteristics on individual and communal levels and the information present on the social network convey actual personality and inclination rather than self-idealization. Homophily is reported to be another important factor which governs the principle of information localization throughout the community. Moreover, status and value homophily play important role in forming individual and collective opinions. These opinions are expressed in free-form keywords on the online social forum eventually collaborating to information organization, retrieval and contextualisation. Facebook and Flickr are prominent platforms for this activity formally termed as folksonomy. However it suffers from issues such as polysemy, ambiguous tags greatly degrade the performance of information retrieval system.

The problem of keyword ambiguity is studied in many interrelated disciplines of computer science such as computational linguistics, artificial intelligence and information systems. Word sense disambiguation (WSD) broadly caters the area of synonymy or ambiguity. Academics have employed supervised and unsupervised methods using statistical, knowledge based and clustered procedures. Although comparison between them is unjustified, the studies using unsupervised methods are more dynamic in identifying different contexts and also reducing human efforts as they do not require a knowledge source. Therefore we can establish from previous work that unsupervised techniques perform better tag disambiguation in Web search sessions. Clusters identify the context and classifiers enable filtered results. We can also conclude that context clustering technique and k-nearest-neighbour algorithm show better results in terms of capturing social contexts as they handle the issue of context duplication and redundancy.

After critical analysis of literature surrounding social network sites and its effectiveness in resolving polysemy in social tagging, we tested our hypothesis. We adapted a comparative evaluation approach to validate our claims. Our methodology included combination of user feedback and experimental comparison of algorithm. As our experiment fundamentally depends on user data, ethics approval was requested from ECS ethics committee. All information and implementation was performed on university virtual machines. In terms of approach, we first ask the user to identify the context of the ambiguous keyword; this gives us a ground truth for evaluating the results. The user's

social network information is used as data source. The disambiguation activity includes identifying the clusters of the queried tag, translate them to classifiers and then compute their similarity with the keywords stored in the database.

Our proposed method involves first retrieving the clusters Flickr pertaining to a tag. These clusters represent contexts in which a tag can be interpreted; however there is level of redundancy in them. For this we make classifiers using the K-Nearest-Neighbour algorithm. The user keywords are then compared to these classifiers to valuate a degree of similarity using the dice co-efficient. The classifiers are arranged in descending orders of similarity and each classifier points to set of images which are ultimately displayed as results. We repeat this process with the keywords of friends to test the existence and effectiveness of homophily in social networks. After conducting this activity for each tag, we finally evaluate our results by calculating the accuracy in compliance to what user has submitted as contexts. We also perform quantitative analysis involving precision and recall of the images retrieved.

In summary, our proposed method for Web search disambiguation is able to successfully order the images according to user's social network information. Our experiment results and manual examination reports an average accuracy of 65%, meaning almost 7 out of 10 images are according to what user indicated and also according to user's SNS profile. One another aspect of this study was to study the effects of homophily and it results have shown that homophily can improve the effectiveness of disambiguating polysemous tags. We also compare our results to one of the similar studies; although our results are not as precise as theirs, we have identified the factors that can help improve our approach.

This piece of research has successfully validated the postulated claims about the effectiveness of social network data in removing tag ambiguity. However we are aware of the limitations of our technical approach in terms of efficiency. Qualitative analysis has revealed some intriguing factors that degrade the performance of the algorithm and we have addressed some of those issues. In future we plan to address these issues. Facebook 'Likes' are a tremendous source of data pertaining to users preferences and they can be used in harvesting collective intelligence. We have only employed the keyword part of the Facebook 'Likes', the JSON format of a 'Like' contains information like date, time, and category. These sub attributes can further increase the process of disambiguation. Google has recently launched its social networking site called Google Plus; they have introduced the concept of 'Sparks' which capture user's interest, activities, hobbies and likes/dislikes. Our approach can be tested on this social network platform to verify the effectiveness of disambiguation with evidence of homophily.

References

Agirre, E. et al., 2006. Two graph-based algorithms for state-of-the-art WSD. *Computational Linguistics*, (July), pp.585-593. Available at: http://portal.acm.org/citation.cfm?doid=1610075.1610157.

Aris, A., Kumar, R. & Mahdian, M., 2008. Influence and Correlation in Social Networks.pdf. In *KDD '08 Proceeding of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*.

Avery, J., 2010. The Democratization of Metadata: Collective Tagging, Folksonomies and Web 2.0. *Library Student Journal*, 5. Available at: http://www.librarystudentjournal.org/index.php/lsj/article/view/135/268.

Back, M.D. et al., 2010. Facebook profiles reflect actual personality, not self-idealization. *Psychological Science*, 21(3), pp.372-374. Available at: http://www.ncbi.nlm.nih.gov/pubmed/20424071.

Barabási, A.-L., 2002. *Linked: The New Science of Networks*, Perseus Publishing. Available at: http://link.aip.org/link/?AJPIAS/71/409/2.

Beer, D. & Burrows, R., 2007. Sociology and, of and in Web 2.0: Some Initial Considerations. *Sociological Research Online*, 12(5), pp.1-15. Available at: http://www.socresonline.org.uk/12/5/17.html.

Beer, D.D., 2008. Social network(ing) sites…revisiting the story so far: A response to danah boyd & Nicole Ellison. *Journal of Computer-Mediated Communication*, 13(2), pp.516-529. Available at: http://doi.wiley.com/10.1111/j.1083-6101.2008.00408.x [Accessed July 4, 2011].

Boyd, D.M. & Ellison, N.B., 2007. Social Network Sites: Definition, History, and Scholarship. *Journal of Computer-Mediated Communication*, 13(1), pp.210-230. Available at: http://doi.wiley.com/10.1111/j.1083-6101.2007.00393.x [Accessed July 15, 2011].

Breslin, J.G., Passant, A. & Decker, S., 2009. *The Social Semantic Web*, Springer-Verlag. Available at: http://www.springerlink.com/index/10.1007/978-3-642-01172-6.

Cashmore, P., 2010. Google's nightmare: Facebook "Like" replaces links. Available at: http://edition.cnn.com/2010/TECH/04/29/cashmore.google.facebook/index.html.

Chi, E.H. & Alto, P., 2008. The Social Web : Research and Opportunities. *Computer*, vol 41, pp.88-91. Available at: http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=4623229&isnumber=4623205.

De Choudhury, M. et al., 2010. "Birds of a Feather": Does User Homophily Impact Information Diffusion in Social Media? *Arxiv preprint arXiv10061702*, p.31. Available at: http://arxiv.org/abs/1006.1702.

Garca-Silva, A. et al., 2009. Preliminary results in tag disambiguation using dbpedia. In *International Conference on Knowledge Capture-Workshop on Collective Knowledge Capturing and Representation, Redondo Beach, CA, USA*. Citeseer. Available at: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.155.2791&rep=rep1&type=pdf [Accessed September 15, 2011].

Golder, S.A. & Huberman, B.A., 2006. Usage patterns of collaborative tagging systems. *Journal of Information Science*, 32(2), pp.198-208. Available at: http://jis.sagepub.com/cgi/doi/10.1177/0165551506062337 [Accessed July 19, 2011].

Gosling, S D, Gaddis, S & Vazire, S, 2007. Personality Impressions Based on Facebook Profiles. *Psychology*, pp.1-4. Available at: http://www.icwsm.org/papers/3--Gosling-Gaddis-Vazire.pdf.

Gosling, Samuel D, Rentfrow, P.J. & Jr, W.B.S., 2003. A very brief measure of the Big-Five personality domains. *Journal of Research in Personality*, 37(6), pp.504-528. Available at: http://linkinghub.elsevier.com/retrieve/pii/S0092656603000461 [Accessed July 21, 2011].

Guy, M. & Tonkin, E., 2006. Folksonomies Tidying up Tags ? *DLib Magazine*, pp.1-17.

Ide, N. & Jean, V., 1997. Word Sense Disambiguation : The State of the Art. *New York*, pp.1-41.

Ireson, N., 2010. Toponym resolution in social media. *Movie*, (1). Available at: http://www.springerlink.com/index/N4Q973263W5007GL.pdf [Accessed September 14, 2011].

Iskold, A., 2007. The New Face of Amazon - Tags, Ajax, Plogs & Wikis. Available at: http://www.readwriteweb.com/archives/amazon_tags_ajax_plogs_wikis.php [Accessed August 2011].

James, K., 2007. Json.Net. Available at: http://james.newtonking.com/projects/json-net.aspx.

Kato, M. et al., 2008. Can Social Tagging Improve Web Image Search? *Web Information Systems Engineering-WISE 2008*, pp.235–249. Available at: http://www.springerlink.com/index/336742L893508V22.pdf [Accessed September 15, 2011].

Lada, A., Orkut, B. & Eytan, A., 2003. A social network caught in the Web. *First Monday*, Vol 6-2, pp.1-22.

Lauw, H. et al., 2010. Homophily in the Digital World: A LiveJournal Case Study. *IEEE Internet Computing*, 14(2), pp.15-23. Available at: http://www.hadylauw.com/ic10.pdf.

Lazarsfeld, P.F. & Merton, R.K., 1954. Friendship as a social process: A substantive and methodological analysis. In M. Berger, T. Abel, & C. H. Page, eds. *Freedom and Control in Modern Society*. Van Nostrand, pp. 18-66. Available at: http://www.questia.com/PM.qst?a=o&docId=23415760.

Lee, K. et al., 2009. Tag Sense Disambiguation for Clarifying the Vocabulary of Social Tags. *2009 International Conference on Computational Science and Engineering*, pp.729-734. Available at: http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=5283425 [Accessed July 19, 2011].

Lin, D., 1998. Automatic retrieval and clustering of similar words. *Proceedings of the 36th annual meeting on Association for Computational Linguistics*, 2, pp.768-774. Available at: http://portal.acm.org/citation.cfm?doid=980691.980696.

Liu, K., Fang, B. & Zhang, W., 2010. Unsupervised Tag Sense Disambiguation in Folksonomies. *Journal of Computers*, 5(11), pp.1715-1722. Available at: http://ojs.academypublisher.com/index.php/jcp/article/view/3035 [Accessed July 19, 2011].

Marlow, C. et al., 2006. HT06, tagging paper, taxonomy, Flickr, academic article, to read U. K. Wiil, P. J. Nürnberg, & J. Rubart, eds. *October*, 27(3), pp.31-40. Available at: http://portal.acm.org/citation.cfm?id=1149949.

McCarthy, D., 2009. Word Sense Disambiguation: An Overview. *Language and Linguistics Compass*, 3(2), pp.537-558. Available at: http://doi.wiley.com/10.1111/j.1749-818X.2009.00131.x.

McPherson, M., Smith-Lovin, L. & Cook, J.M., 2001. Birds of a Feather: Homophily in Social Networks. *Annual Review of Sociology*, 27(1), pp.415-444. Available at: http://www.annualreviews.org/doi/abs/10.1146/annurev.soc.27.1.415.

Mika, P., 2007. Ontologies are us: A unified model of social networks and semantics. *Web Semantics: Science, Services and Agents on the World Wide Web*, 5(1), pp.5-15. Available at: http://linkinghub.elsevier.com/retrieve/pii/S1570826806000552.

Milgram, S., 1967. The small world problem. *Psychology Today*, 2(1), pp.60-67. Available at: http://measure.igpp.ucla.edu/GK12-SEE-LA/Lesson_Files_09/Tina_Wey/TW_social_networks_Milgram_1967_small_world_problem.pdf.

Pantel, P. & Lin, D., 2002. Discovering word senses from text. *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining KDD 02*, 41, p.613. Available at: http://portal.acm.org/citation.cfm?doid=775047.775138.

Papagelis, M. & Murdock, V., 2011. Individual Behavior and Social Influence in Online Social Systems. In *HT '11 Proceedings of the 22nd ACM conference on Hypertext and hypermedia*. pp. 241-250.

Peterson, E., 2006. Beneath the metadata: Some philosophical problems with folksonomy. *DLib Magazine*, 12(11), pp.1-6. Available at: http://www.dlib.org/dlib/november06/peterson/11peterson.html.

Purandare, A. & Pedersen, T., 2004. Word sense discrimination by clustering contexts in vector and similarity spaces. *Proceedings of the Conference on Computational Natural*

Language Learning, pp.41–48. Available at: http://acl.ldc.upenn.edu/hlt-naacl2004/conll04/pdf/purandare.pdf.

Rainie, L., 2007. *28% of online americans have used the internet to tag content.Tech- nical report, Pew Internet and American Life Project*,

Rosen, C., 2007. Virtual Friendship and the New Narcissism. *The New Atlantis*, pp.15-31.

Schütze, H., 1998. Automatic Word Sense Discrimination. *Computational Linguistics*, 24(1), pp.97-123.

Singla, P. & Richardson, M., 2008. Yes , There is a Correlation - From Social Networks to Personal Behavior on the Web. *Computing*, pp.655-664.

Sun, J., 2010. Flickr.Net API. Available at: http://www.codeplex.com/wikipage?ProjectName=FlickrNet.

Thelwall, M., 2009. Homophily in MySpace. *Journal of the American Society for Information Science and Technology*, 60(2), pp.219-231. Available at: http://onlinelibrary.wiley.com/doi/10.1002/asi.20978/full.

Utz, S., 2010. Show me your friends and I will tell you what type of person you are: How one's profile, number of friends, and type of friends influence impression formation on social network sites. *Journal of Computer-Mediated Communication*, 15(2), pp.314-335. Available at: http://doi.wiley.com/10.1111/j.1083-6101.2010.01522.x [Accessed June 17, 2011].

Vanderwal, T., 2007. Folksonomy Coinage and Definition. Available at: http://www.vanderwal.net/folksonomy.html.

Veronis, J., 2004. HyperLex: lexical cartography for information retrieval. *Computer Speech Language*, 18(3), pp.223-252. Available at: http://apps.isiknowledge.com.libproxy.unm.edu/full_record.do?product=WOS&colname=WOS&search_mode=RelatedRecords&qid=858&SID=1BFE94Ekeg2KHDJkJJ8&page=3&doc=27.

Wellman, B., 1988. Structural analysis: from method and metaphor to theory and substance. In B. Wellman & S. D. Berkowitz, eds. *Social Structures A Network Approach*. Cambridge University Press, pp. 19-61. Available at: http://scholar.google.com/scholar?hl=en&btnG=Search&q=intitle:Structural+analysis:+From+method+and+metaphor+to+theory+and+substance#0.

Yeung, C.-man A., 2009. From User Behaviours to Collective Semantics by October 2009. *Discovery*, (October).

Yeung, C.-man A., Gibbins, N. & Shadbolt, N., 2008. A k-Nearest-Neighbour Method for Classifying Web Search Results with Data in Folksonomies. In *2008 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology*. Sydney, Australia: IEEE, pp. 70-76. Available at: http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=4740428.

Yeung, C.-man A., Gibbins, N. & Shadbolt, N., 2007. Tag Meaning Disambiguation through Analysis of Tripartite Structure of Folksonomies. *2007 IEEE/WIC/ACM International Conferences on Web Intelligence and Intelligent Agent Technology - Workshops*, pp.3-6. Available at: http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=4427527.

Appendices

Appendix A: Project Brief

2011 MSc Project Brief

Name	**Syed Sumair Qasim**	**ID no**	**24368032**	**Email**	**ssq1g10@ecs.soton.ac.uk**

Supervisor	**Dr. Mark J Weal**	**2nd Examiner**	
Co-supervisor			

Please note: if you need advice or support at any stage of your project and you can not contact your supervisor, you should contact your co-supervisor or your second examiner. If you can not contact any of these, and it is important, then contact your Course Leader or the MSc director.

Project title *(this may change)*	**Tag Disambiguation based on information drawn on Social Networks**

Description of project

In today's world of Web 2.0 where users are responsible for generating and sharing content on the global information space, they are also accountable for replicating the information. This particular case can be observed in social tagging or collaborative tagging systems which allow users to make use of keywords(tags) to describe publically available Web content and this entire exercise constitutes to a concept called Folksonomy as described by T.V. Wal. One major issue encountered in this activity is the ambiguity of terms. One word can be understood with different meanings as the context is not defined.

Several studies have employed different approaches to get hold around the contextualisation of tags which includes applying data mining techniques on the large data set to form clusters. Au Yeung et al researched on the associations of a single tag with other tags, users and documents in a folksonomy. However there is little work done on resolving tag meaning with respect to social networks which has been the most widely used service of the Web 2.0. Therefore, we propose that by making use of information available in social networks we can contextualise the keyword/tag with more precision as compared to the techniques which don't use the social context. We believe that this would also help improve interest matching in social networks. One possible scenario would be an application aggregating the interests from a Facebook user's profile and populating the clusters based on gathered information. When queried, it will return the results filtered according to the user's profile. One another aspect that we will also handle is the inclusion of interests of the user's friends (homophily), which can be useful when there is insufficient information gathered from user's profile.

We will go about doing this by conducting initially two experiments: one without using the social networks information (traditional method) and the other method using the social contexts. It is to note that both approaches will have the same dataset as we will be evaluating the results based on the comparison of these two approaches. The plan includes initial few weeks for gathering the relevant literature and finalizing the method (algorithm) for comparison of query with user's keywords. This would follow the designing of the framework that would conduct experiments from both aspects (social and traditional).

We have reserved almost one month for development of the application that would test the dataset against the aggregated user data. The results of the testing would be further evaluated to prove our hypothesis. All of these activities will be documented in the report simultaneously with proper citations to contributors. First draft will be made available in the first week of September to allow time for corrections and amendments. A more generic view of the plan can be seen in the following Gantt chart.

Does your project involve laboratory work?	**Does your project involve human subjects?**
NO	**YES and I am working on an ethics application**

Appendix B: Original Gantt Chart

Week # ***week beginning :***	**1** **13/6**	**2** **20/6**	**3** **27/6**	**4** **4/7**	**5** **11/7**	**6** **18/7**	**7** **25/7**	**8** **1/8**	**9** **8/8**	**10** **15/8**	**11** **22/8**	**12** **29/8**	**13** **5/9**	**14** **12/9**	**15** **19/9**
Activities and milestones															
Topic Selection															
Literature Search and Review															
Design and Development															
Testing and Evaluation															
Writing-up															
Milestone – demonstrate to sup/examiner												*			
Milestone – dissertation draft complete													*		
Final corrections															
Milestone – Hand-in															*

Appendix C: ECS Ethics Form

School of Electronics and Computer Science

School Ethics Committee

Application for Expedited Review (subject data limited)

Please refer to the definitions of the terms shown in **bold**, and the instructions which follow the form. The form is to be completed by the **investigator**(s). The **project** is eligible for expedited review if it is able to provide the answers not greyed out to questions 1-8, otherwise the proposal requires normal review.

OFFICE USE ONLY
Reference number:
ES ___/___/____
[yy/mm/nnn]

Name of **investigator**(s): Syed Sumair Qasim	Date of application: **Signature**(s):
Name of supervisor(s): MARK WEAL	**Signature**(s): (for MARK WEAL)
Title of **project**: Tag disambiguation based social network information	

If the **project** is **invasive**, involves **minors**, animals, or coercion, please refer to your Group Representative on the School Ethics Committee.

		Yes	No
1	Does the **project** involve human **participants**? If 'No', approval for the **study** may be dealt with by Chair's Action.	Yes	
2	Is the **risk** of **harm** to any **participant insignificant**?	Yes	
3	Will every **participant** be able to withdraw from the **project** at any time and for any reason?	Yes	
4	Will every **participant** be informed of the true purpose of the **project** before the **project** begins?	Yes	
5	Does the **project** involve deception of any **participant**?		No
6	Does the **project** involve **sensitive data**?		No
7	Is the **project intrusive**?		No
8	If the **project** involves **subject data**, will all **participants** be students of the University? If the project involves **subject data** in respect of one or more **participants** who are NOT students of the University, this is not the correct form of submission for ethical review. Please use the 'E' form instead.	Yes	
9	Does the **project** involve **inducement** to any **participant**? If 'Yes', explain the nature of the **inducement** (eg course credit) in an attached paragraph.		No

Attach a **description** of the **project**. Attach copies of any questionnaires or other **project** instruments. Version number every document.

Name of ECS SEC approving member:	Date of approval:
	Signature:

Appendix D: Results of Manual Examination (Selected Tags)

Architecture

User	Context	List 1: Simple Query	List 2: User's 'Likes'	List 3 Friends' 'Likes'
1.	Software	0.2	0.4	0.5
2.	Physical Structure	0.5	0.6	0.8
3.	Physical Structure	0.5	0.6	0.7
4.	Software	0.1	0.6	0.5
5.	Physical Structure	0.6	0.8	0.5
6.	Physical Structure	0.5	0.6	0.5
7.	Physical Structure	0.7	0.8	0.4
8.	Software	0.3	0.6	0.6
9.	Software	0.3	0.7	0.5
10.	Physical Structure	0.7	0.9	0.4
Average		0.44	0.66	0.53

Soap

User	Context	List 1: Simple Query	List 2: User's 'Likes'	List 3 Friends' 'Likes'
1.	Cleaning Agent	0.5	0.6	0.5
2.	Web Service	0.5	0.2	0.8
3.	Cleaning Agent	0.5	0.6	0.7
4.	Web Service	0.1	0.2	0.5
5.	Web Service	0.2	0.2	0.5
6.	Cleaning Agent	0.5	0.6	0.5
7.	Web Service	0.3	0.3	0.4
8.	Web Service	0.3	0.2	0.6
9.	Cleaning Agent	0.5	0.7	0.5
10.	Cleaning Agent	0.7	0.6	0.4
Average		0.41	0.4	0.53

Appendix E: Source Code