UNIVERSITY OF SOUTHAMPTON

From User Behaviours to Collective Semantics

by

Ching Man Au Yeung

A thesis submitted in partial fulfillment for the degree of Doctor of Philosophy

in the
Faculty of Engineering and Applied Science
Department of Electronics and Computer Science

October 2009

UNIVERSITY OF SOUTHAMPTON

ABSTRACT

FACULTY OF ENGINEERING AND APPLIED SCIENCE
DEPARTMENT OF ELECTRONICS AND COMPUTER SCIENCE

Doctor of Philosophy

by Ching Man Au Yeung

The World Wide Web has developed into an important platform for social interactions with the rise of social networking applications of different kinds. Collaborative tagging systems, as prominent examples of these applications, allow users to share their resources and to interact with each other. By assigning tags to resources on the Web in a collaborative manner, users contribute to the emergence of complex networks now commonly known as folksonomies, in which users, documents and tags are interconnected with each other. To reveal the implicit semantics of entities involved in a folksonomy, one requires an understanding of the characteristics of the collective behaviours that create these interconnections. This thesis studies how user behaviours in collaborative tagging systems can be analysed to acquire a better understanding of the collective semantics of entities in folksonomies. We approach this problem from three different but closely related perspectives. Firstly, we study how tags are used by users and how their different intended meanings can be identified. Secondly, we develop a method for assessing the expertise of users and quality of documents in folksonomies by introducing the notion of implicit endorsement. Finally, we study the relations between documents induced from collaborative tagging and compare them with existing hyperlinks between Web documents. We show that, in each of these scenarios, it is crucial to consider the collective behaviours of the users and the social contexts in order to understand the characteristics of the entities. This project can be considered as a case study of the Social Web, the research outcomes of which can be easily generalised to many other social networking applications. It also fits into the larger framework for understanding the Web set out by the emerging interdisciplinary field of Web Science, as the work involves analyses of the interactions and behaviour of Web users in order to understand how we can improve existing systems and facilitate information sharing and retrieval on the Web.

Contents

List of Figures

List of Tables

Declaration of Authorship

I, Ching Man Au Yeung, declare that the thesis entitled *User Behaviours and Collective Semantics* and the work presented in the thesis are both my own, and have been generated by me as the result of my own original research. I confirm that:

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

Collaborative Work

The majority of the research work described in Chapter 5, 'Implicit Endorsement and Expertise Ranking', was collaboratively conducted with Michael Noll from the University of Potsdam. We worked together on formulating the method for measuring user expertise, designing the SPEAR algorithm, and implementing the simulation experiments for evaluating SPEAR. Michael was primarily responsible for setting up the experimentation environment and coding, while I was mainly involved in the execution and customisation of the simulations. I extended the discussion of implicit endorsement in this thesis and conducted a further study on the credit score functions. For other parts of the work we collaborated closely with each other.

List of Publications

- Ching Man Au Yeung, Nicholas Gibbins, and Nigel Shadbolt. Mutual contextualization in tripartite graphs of folksonomies. In *ISWC 2007: The 6th International Semantic Web Conference, Pusan, South Korea, 11-15 November*, pages 966–970. Springer, 2007.

- Ching Man Au Yeung, Nicholas Gibbins, and Nigel Shadbolt. Understanding the semantics of ambiguous tags in folksonomies. In *The International Workshop on Emergent Semantics and Ontology Evolution (ESOE2007) at ISWC/ASWC 2007, Pusan, South Korea, 11-15 November*, pages 108–121, 2007.

- Ching Man Au Yeung, Nicholas Gibbins, and Nigel Shadbolt.Tag meaning disambiguation through analysis of tripartite structure of folksonomies. In *The 2007 IEEE / WIC / ACM International Conference on Web Intelligence and Intelligent Agent Technology - Workshops, Silicon Valley, CA, USA, 2-5 November*, pages 3–6, 2007.

- Ching Man Au Yeung, Nicholas Gibbins, and Nigel Shadbolt. Web search disambiguation by collaborative tagging. In *Proceedings of the Workshop on Exploiting Semantic Annotations in Information Retrieval (ESAIR 2008), co-located with ECIR 2008, Glasgow, United Kingdom, 31 March 2008*, pages 48–61, 2008.

- Ching Man Au Yeung, Nicholas Gibbins, and Nigel Shadbolt. A study of user profile generation from folksonomies. In *Proceedings of the Workshop on Social Web and Knowledge Management (SWKM2008) at WWW2008, Beijing, China, 21-25 April*, pages 1–8, 2008.

- Ching Man Au Yeung, Nicholas Gibbins, and Nigel Shadbolt. Collective user behaviour and tag contextualisation in folksonomies. In *The 2008 IEEE / WIC / ACM International Conference on Web Intelligence and Intelligent Agent Technology - Workshops, Sydney, Australia, 9-12 December*, pages 659–662. IEEE Computer Society Press, 2008.

- Ching Man Au Yeung, Nicholas Gibbins, and Nigel Shadbolt. Discovering and modelling multiple interests of users in collaborative tagging systems. In *The 2008 IEEE / WIC / ACM International Conference on Web Intelligence and Intelligent Agent Technology - Workshops, Sydney, Australia, 9-12 December*, pages 115–118. IEEE Computer Society Press, 2008.

- Ching Man Au Yeung, Nicholas Gibbins, and Nigel Shadbolt. A k-nearest-neighbour method for classifying web search results with data in folksonomies. In *Proceedings of the 2008 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology, Sydney, Australia, 9-12 December*, pages 70–76. IEEE Computer Society Press, 2008.

- Ching Man Au Yeung, Michael G. Noll, Nicholas Gibbins, Christoph Meinel, and Nigel Shadbolt. On measuring expertise in collaborative tagging systems. In *WebSci'09: Web Science Conference 2009 - Society On-Line*, 2009.

- Ching Man Au Yeung, Nicholas Gibbins, and Nigel Shadbolt. Contextualising tags in collaborative tagging systems. In *Proceedings of the 20th ACM Conference on Hypertext and Hypermedia, 29 June - 1 July, 2009, Turino, Italy*, pages 251–260. ACM, 2009.

- Michael G. Noll, Ching Man Au Yeung, Nicholas Gibbins, Christoph Meinel, and Nigel Shadbolt. Telling experts from spammers: Expertise ranking in folksonomies. In *Proceedings of the 32nd Annual ACM SIGIR Conference, 19-23 July, 2009, Boston, MA, USA*, pages 612–619. ACM, 2009.

- Ching Man Au Yeung, Nicholas Gibbins, and Nigel Shadbolt. User-induced links in collaborative tagging systems. In *Proceedings of the 18th ACM Conference on Information and Knowledge Management, 2-6 November, 2009, Hong Kong*. ACM, 2009.

- Ching Man Au Yeung, Nicholas Gibbins, and Nigel Shadbolt. Multiple interests of users in collaborative tagging systems. In Ricardo Baeza-Yates and Irwin King, editors, *Weaving Services and People on the World Wide Web*, pages 255–274. Springer, 2009.

Ching Man Au Yeung
29 September 2009

Acknowledgements

I would first like to express my gratitude towards my supervisors Dr. Nicholas Gibbins and Prof. Nigel Shadbolt. They gave me a lot of invaluable advice and suggestions over the past three years. They also inspired me in many occasions, which led to many of the ideas presented in this thesis. I also thank Prof. Wendy Hall and Prof. Alun Preece for their comments and suggestions that helped me to improve this thesis.

I also wish to thank the Drs Richard Charles and Esther Yewpick Lee Charitable Foundation for its exceedingly generous financial support made through the R C Lee Centenary Scholarship. I own a great debt of gratitude to Dr. Deanna Rudgard, the chairperson of the foundation, who has been caring and supportive ever since I got to know her.

In addition, grateful thanks are due to Michael G. Noll, a PhD candidate at the Hasso Plattner Institut, University of Potsdam. We got to know each other in 2007 at the International Semantic Web Conference. Since then we have been collaborating on different projects. Some sections in this thesis greatly benefited from Michael's contribution and discussion with him.

My research work was conducted in the School of Electronics and Computer Science, which provided great support to my research activities and participation in international conferences. I also benefited from discussion, comments and suggestions provided by fellow postgraduate students. My visit to the Massachusetts Institute of Technology in Boston in Fall 2008 also allowed me to refine my ideas through discussion with researchers and students at the Decentralized Information Group, CSAIL.

Last but not least, I thank my family for their unconditional support and encouragement throughout the years of my studies.

Chapter 1

Introduction

The World Wide Web, or simply the Web, has experienced dramatic growth since its inception in the 1990s. It was estimated that there were already over 11.5 billion Web pages on the Web as of the end of January 2005 (Gulli and Signorini, 2005).[1] The Web provides the infrastructure for people to publish information online and to link related resources with each other by using hyperlinks, resulting in a huge network of information that has revolutionised the way people disseminate and exchange information. Nowadays, for any person with a personal computer and an Internet connection, obtaining information has never been so quick and convenient. For example, details of public services can be found on the Web sites of governments, timetables and maps of public transports can be downloaded from Web sites of corresponding companies, news from around the world is only a few clicks away on mass media Web sites, and even learning a foreign language is made possible and much easier by all the textual and audiovisual resources available on the Web.

Besides being a medium for disseminating information, the Web in recent years has also become an important platform of social interactions. Along with the trend of Web 2.0, which is now also often referred to as the Social Web (Chi, 2008), we see a large number of Web sites offering different kinds of tools for users to interact with each other and to establish their social networks online. Weblogs, or nowadays more commonly known as blogs, provide general Web users the convenience of publishing their ideas and thoughts on the Web without the need to worry about the cost of publishing or the trouble of learning to use even the most

[1] The Web site WorldWideWebSize.com reported on 26 April 2009 that the size of the indexed Web was estimated to be at least 23.46 billion pages.

user-friendly Web design software. Wikipedia[2] represents an excellent example of how collaborative efforts of Web users can be transformed into a multilingual encyclopedia on the Web. Facebook[3], MySpace[4], Mixi[5], Orkut[6] to name but a few social networking sites allow users to share their interests and to keep track of what their friends are doing. And then there are user-driven knowledge sharing sites such as Yahoo! Answers that allow users to ask and answer questions (Adamic et al., 2008).[7] Finally, there are the tags contributed by Web users: freely-chosen descriptive keywords that are used by Web users to describe, organise, share and retrieve Web resources in collaborative tagging systems, such as Delicious[8] and Bibsonomy[9].

In fact, tagging is one of the most prominent features on the Social Web. In the most general sense, such functionality enables Web users to express their viewpoints on how the Web should be organised. They do this by choosing tags they like to describe the resources that they find interesting and useful. In addition, collaborative tagging represents a very general form of the new type of interactions promoted by the Social Web. These interactions involve three entities, namely Web users, Web documents or resources, and tags. We believe that these are the three most important entities on the Social Web. Their interactions should be studied in order to acquire a better understanding of the Social Web, so as to improve existing as well as to devise new Social Web applications to facilitate interactions between Web users.

Of course, there is no doubt that collaborative tagging systems represent only a small subset of the large number of social Web sites and applications available nowadays. Also, these systems do not necessarily provide users with functionalities that allow them to create explicit social networks online. However, collaborative tagging is almost everywhere on the Web. It finds its way into different applications where annotation and organisation of information are needed. In fact, many online interactions nowadays can be considered as a special case of collaborative tagging, as they mostly involve users expressing, either implicitly or explicitly, their preferences and opinions on online resources. For example, Krause et al. (2008) show that a search engine query log with data of search keywords and

[2]Wikipedia: `http://www.wikipedia.org/`
[3]Facebook: `http://www.facebook.com/`
[4]MySapce: `http://www.myspace.com/`
[5]`Mixi:http://www.mixi.jp/`
[6]Orkut: `http://www.orkut.com/`
[7]Yahoo! Answers: `http://answers.yahoo.com/`
[8]Delicious: `http://delicious.com/`
[9]BibSonomy: `http://www.bibsonomy.org/`

browsing history can be modelled as a folksonomy, a data structure commonly found in collaborative tagging. In addition, instead of explicit social ties, we are more interested in the implicit relations between entities on the Web, which can be very valuable for understanding the qualities and characteristics of not only users but also keywords and documents.

One intriguing feature of the phenomenon of collaborative tagging is that, like the Web, it is based on a very simple idea of allowing users to assign descriptive keywords to online resources, but evolves into a huge and complex structure as it becomes more and more popular. In the case of the Web, the simple idea is the act of creating a hyperlink from one hypertext document to another. Out of this simple idea, we now have a huge and complex network of interlinking documents that requires much more effort and more sophisticated techniques to understand (Hendler et al., 2008). In the case of tagging, we have a folksonomy, a complex network structure of users, tags and documents. The associations between these entities encoded in a folksonomy convey a lot of implicit information about their characteristics, such as the meaning of the tags, the trustworthiness and expertise of the users, the topics and quality of the documents, in a way same as how the link structure of the Web convey information about the flow of information, the popularity, the usefulness and the interrelations of the hypertext documents.

This thesis sets out to explore the different characteristics of the entities in a folksonomy. More specifically, we aim to understand what we call the ***collective semantics*** embedded in the associations resulted from the collective user behaviours in collaborative tagging, and to investigate how such understanding facilitate further social interactions and information organisation. By collective semantics, we refer to the meanings, interpretations, and/or qualities acquired by an entity in a folksonomy through the process of collaborative tagging. Collective semantics emerges as a result of the collective behaviour guided by a certain set of rules, regardless of the exact intents of individual users. The notion of complex network is important in studying the collective semantics of entities in a folksonomy because the associations these entities are involved in sheds light on how they are collectively interpreted by the users.

In the following sections, we continue this chapter by presenting a brief introduction to the Social Web and collaborative tagging, followed by a discussion of the notion of collective semantics. We then specify the research questions that motivate this thesis and highlight the contributions we have made to this area.

1.1 From the Web to the Social Web

The Web has rapidly become an important medium of information dissemination since it was first invented in the early 1990s. It is a universally accessible and enormous network of hypertext documents (Web pages) involving not only textual content but also images, sounds, videos and other types of multimedia content. Any person with a personal computer and access to the Internet can create some Web pages and upload them to a Web server, and then the content of these Web pages will be available to other people who also have access to the Web.

There should be little doubt that the Web has not only been a technological tool but also a medium of social interaction. Users with enough knowledge of online publishing—such as a basic understanding of HTML or of one of those WYSIWYG (What-You-See-Is-What-You-Get) Web authoring applications and file uploading applications—can readily present something on their homepages, which can contain hyperlinks to other documents on the Web. Hyperlinks act as something that binds together documents that are created by different authors. They are also used by Web authors as a form of endorsement or recommendation. Creating a hyperlink from my homepage to your homepage can be considered as a way of expressing recognition. In the early days of the Web, Web sites or personal homepages would organise themselves around something called Web rings, which were collections of pages of specific themes. A member of a Web ring would place a navigation banner in his/her homepage, thus connecting his/her page to those of the rest of the members.

However, in the early days of the Web, there was still a certain level of 'entry requirement' for someone who would like to publish something online. Having a computer and an internet connection was not enough to do so, if one did not have the necessary technical knowhow. Hence, it was quite natural that users of the Web were divided into those who provided information and those who consumed the information, as Resnick (1998) remarks that '[t]he real class division in Cyberspace is between the Webmasters and the Surfers'. About.com can probably be considered as an exemplar of content providers in the early days of the Web. The Web site provides information and resources to Web users on a wide range of topics, ranging from practical skills such as fishing and gardening to theories in various disciplines in natural science.[10] Content is compiled with the help of a team of experts in the corresponding areas. It represents a typical example of the author-versus-reader paradigm of the early Web.

[10] About.com: `http://www.about.com/`

We have seen a significant change of the Web—the so-called Web 2.0 or Social Web—regarding this author-versus-reader paradigm in recent years. With the rises of Weblogs (blogs) and wikis, social bookmarking services, collaborative filtering sites, image and video sharing sites, and social networking sites, Web users nowadays find it much easier to contribute content to the Web. It is therefore also much easier for them to alternate between the roles of authors and readers. For example, blogs allow Web users to publish their thoughts, ideas, information, or even research findings on the Web with little effort. While they can run a blog application on a server by themselves, there are also plenty of Web sites out there where they can register for a blog for free and start writing, which in general involves only filling out a Web-based form. Providing similarly easy-to-use editing functions, wikis allow users to create and edit content of a page in a collaborative manner. Their flexibility and usefulness are clearly demonstrated in the popularity of Wikipedia, a free online encyclopedia created and edited collaboratively by ordinary Web users. Of course, such so-called user-generated content is not limited to textual information. Other forms of resources, such as digital images and video clips, can be found on popular social Web sites like Flickr[11] and YouTube[12].

However, the emergence of this kind of social Web applications does not only encourage Web users to contribute information, but also information about information. Many social Web applications allow users to express their opinions on how online resources should be organised. Collaborative tagging is a prominent example of this kind. Web sites such as Delicious, BibSonomy, LibraryThing[13], Last.fm[14] and CiteULike[15] allow users to assign freely-chosen descriptive keywords, commonly known as tags, to shared resources for the purpose of organisation, sharing, and facilitation of future retrieval. In the past, an author is supposed to be responsible for providing metadata such as the topic and the keywords of his document to facilitate other users to judge whether it is relevant to their needs. Collaborative tagging systems, on the other hand, allow the readers of the documents to decide how they should be described, annotated and categorised. There are also related social applications, such as Digg[16] and Reddit[17], in which users can vote, rate or comment on resources available on the Web, thus allowing users to determine which resources are useful, relevant, interesting and/or worth reading.

[11] Flickr: `http://www.flickr.com/`
[12] YouTube: `http://www.youtube.com/`
[13] LibraryThing: `http://www.librarything.com/`
[14] Last.fm: `http://www.last.fm/`
[15] CiteULike: `http://www.citeulike.org/`
[16] Digg: `http://digg.com/`
[17] Reddit:`http://www.reddit.com/`

While it is suggested that in terms of technology the Web 2.0/Social Web bears no significant difference from the original Web, it is without doubt that there is a change in the form of participation of Web users.[18] Thanks to the numerous social Web sites, social interactions on the Web have never been so ubiquitous before. In addition to the major services provided by these Web sites, many of them also allow Web developers to access their data through APIs (application programming interfaces) programmatically for the purpose of developing new applications that use the data or add value to the data by combining it with that from other Web sites. In summary, it can be seen that overall the Social Web not only promotes user-generated content but also encourages users to add values to online resources in a collaborative manner through different kinds of social interactions on the Web.

1.2 Social and Collaborative Tagging

As we have just mentioned in the previous section, a very prominent feature of the Social Web is the prevalent functionality of allowing users to annotate and rate online content in many social Web sites. Tagging refers to the act of assigning freely-chosen descriptive keywords to online content by ordinary Web users. Originally promoted by social bookmarking sites such as Delicious back in 2005, tagging has gained its popularity ever since, finding its way into organisation and sharing of photos (e.g. Flickr), books (e.g. LibraryThing), academic references (e.g. Bibsonomy and Connotea) and news (e.g. Reddit and Digg). Other Web sites have also started to allow users to assign tags to their resources to solicit user annotations. For example, in 2006 Amazon started to allow users to assign tags to books and other products sold on their Web site (Iskold, 2007). Tagging also finds its use in video sharing sites such as Youtube for indexing and retrieval purposes. According to a report on the usage of collaborative tagging published in 2007 (Rainie, 2007), about 28% of American Internet users have engaged in some forms of tagging activities.

While tagging in its simplest form is no more than manual subject indexing (a topic which we will briefly discuss in Chapter 2), in the context of the Social Web it is particularly interesting because its nature is usually social and collaborative. In many popular tagging systems, users do not stop assigning tags to online resources even when some other users have already assigned some tags. Instead,

[18] Tim Berners-Lee remarked in an interview that Web 2.0 'means using the standards which have been produced by all these people working on Web 1.0'. The script can be accessed at `http://www.ibm.com/developerworks/podcast/dwi/cm-int082206txt.html`.

every user can assign their own tags to the resources they found interesting. In other words, every user are allowed to express how they think a resource on the Web should be described, categorised and retrieved. Exemplars of these systems are Delicious, LibraryThing, BibSonomy and CiteULike. These systems aggregate tags contributed by the users, and together these associations between users, resources and tags form a tripartite structure now commonly known as folksonomy.

Folksonomies represent a democratic way of annotating documents on the Web. Unlike metadata provided by the authors of Web documents, a folksonomy reflects the viewpoints of the readers of these documents on how they should be indexed and described. Folksonomies are also believed to be able to provide categorisation schemes that are more flexible, dynamic and up-to-date. One thing special about folksonomies, in contrast to more formal categorisation schemes such as taxonomies and ontologies, is that they are horizontal but not hierarchical. No keywords are subordinate or superordinate of some other tags (except in some systems such as BibSonomy which allows subsumption relations to be specified among tags). In other words, there is no rigid structure that prohibits users from using tags in their own ways.

Tags contributed by users act as intermediates that bring users of similar interests and documents of similar topics together. By following a particular tag, we can identify users who have used the tag frequently. We can also identify documents that have been assigned the tag frequently. In other words, the collective user behaviour in collaborative tagging has generated a lot of new (implicit) links between not only Web documents but also users. Even tags of similar topics are found to be associated with each other because they have been used together very frequently. As a result, collaborative tagging offers users not only a new way of organising and retrieving resources on the Web, but also a new means of exploring the Web by following the links generated by the users themselves. Mathes (2004) notes that serendipity is one of the strengths of folksonomies. They benefit users when they are not looking for specific answers, and when they want to explore the Web in the hope that they will find something interesting.

Collaborative tagging attracts the attention of researchers because it provides a lot of opportunities to extend existing research in such areas as information retrieval and Web search (Heymann et al., 2008a; Noll and Meinel, 2007a; Yanbe et al., 2007), computational linguistics, (Cattuto et al., 2008b), recommendation systems (Niwa et al., 2006; Shepitsen et al., 2008), and the Semantic Web (Specia and Motta, 2007; Van Damme et al., 2007). Folksonomies by themselves are

of particular interest to researchers who study large scale annotation of shared resources. This is because it has been rather difficult to entice Web users to annotate resources for organisational and retrieval purposes before the existence of collaborative tagging systems. In addition, collaborative tagging systems provide a huge amount of data of social interactions on the Web, which also helps to promote studies of user behaviour and social networks in these systems (Golder and Huberman, 2006; Halpin et al., 2007; Mika, 2007), especially when the data in these systems are publicly available and can be collected relatively easily for conducting experiments and analyses. A thorough review of the state-of-the-art studies and analysis of collaborative tagging will be presented in Chapter 2.

Weinberger (2007) describes tagging as an exemplar of the 'third order of order'. In the first order, we deal with things themselves, i.e. to organise the things physically by putting them in different places. In the second order, we come up with catalogues that contain information about the things, i.e. to use metadata to organise the things. In the third order, everything is digitised and things can be sorted at the time when they are retrieved. He also describes this third order as an externalisation of meaning. Things are no longer given any definition in this order. Instead, the meaning of a thing is embedded in the associations and relationships established by whoever interested in them. In fact, this notion of 'meaning' does not only apply to words or tags, but can be extended to described the collective semantics of any entities involved in a folksonomy, including not only the tags but also the users and the documents. In the following section, we will focus on this notion of collective semantics and its relation with user behaviour and its importance in collaborative tagging systems.

1.3 Collective Semantics

Collaborative tagging represents a fundamental form of interactions between three major types of entities on the Web, namely Web users, keywords and Web documents (which include Web pages, images, videos and any other form of online resources).[19] These three entities are the three most important types of entities that characterise the Social Web. While it is more than obvious that users and documents are essential in this setting, keywords are equally indispensable because they represent the medium through which users express their opinions about the

[19]When we say that a user is 'on the Web', we actually refer to the identity of the user on the Web, such as the user name of this user in a particular system with which his/her activities are associated.

documents they have access to on the Web. Many different forms of interactions on the Social Web can be modelled as interactions of these three different types of entities (Krause et al., 2008).

In a collaborative tagging system, the users' activities of using keywords to annotate, describe and categorise documents create a tripartite network structure. Associations (and very often their strengths) between the three types of entities are embedded in such a network structure. These associations, however, are not arbitrary generated. Instead, they are the results of the collective user behaviour observed in the collaborative tagging process. The act of a user assigning a tag to a document involves a certain context, a certain purpose and is usually influenced by the choices of other users (Suchanek et al., 2008). Hence, how different entities in a folksonomy are associated with each other actually reflects something about the meanings, qualities or in general the characteristics of the entities themselves. To develop a framework to study this aspect of the Social Web, we introduce the notion of ***collective semantics***. When we use the term 'semantics', we do not refer to its meanings as defined in linguistics or computer science, in which it is used to refer to the meanings of words or symbols. Instead, we use 'collective semantics' to describe a wider range of characteristics of the tags, users and documents observed as a result of the collective user behaviours on the Web.

An obvious example of such kind of collective semantics stems from co-occurrence analysis of tags. Tags that are frequently used together by the same users or on the same documents can usually be considered as semantically related to each other. For example, the tag `css` is found to be used together with the tag `webdesign` very frequently in Delicious. Based on this idea, Flickr provides clusters of tags that are usually used together on the same photos such that users can explore sets of photos that correspond to different sub-topics of a particular tag (Moëllic et al., 2008). However, by collective semantics, we do not only mean the semantic relations between tags discovered by statistical analysis of co-occurrence. Instead, we focus on the more general characteristics or qualities of the entities involved. These include for example the meanings of a tag in different contexts and among different group of users, the trustworthiness or the expertise of a user, and the relations between documents addressing similar topics from the perspectives of the users. We believe such semantics of the entities can be understood by analysing the complex network structures that have emerged out of the collective behaviour in collaborative tagging.

Similar ideas of analysing collective user behaviour for the emergence of im-

plicit semantics have been discussed under such phrases as collective intelligence (Weiss, 2005; Surowiecki, 2004; Szuba, 2001) and more recently emergent semantics (Aberer et al., 2004; Herschel et al., 2008; Staab, 2002). However, there are differences between these and collective semantics. Collective intelligence focuses on how the aggregation of decisions of individuals produce better answers to questions than any of the individuals does. Examples of collective intelligence can be found in prediction markets and development of open source software. The notion of collective intelligence usually involves a particular task to be solved by a group of individuals, who are in general aware of this. Emergent semantics, on the other hand, is concerned with the semantic interoperability that emerges from agreements among a large number of agents on how certain objects should be interpreted within a certain context (Aberer et al., 2004). Its final product is usually a schema or a specification of objects of a certain domain, it represents the consensus arrived through negotiation among the agents for the purpose of interoperability. Collective semantics is different from the above two concepts in that it is concerned with the semantics that arises from the collective behaviour of users who are not necessarily aware of any particular global tasks. Users assign tags to documents because they want to organise and share, but not because they want to come up with a consensus on the meanings of a tag, or with a ranking of users according to their trustworthiness. Another difference is that it is concerned with a broader range of qualities of the entities involved, instead of only the semantics required for interoperability as in the case of emergent semantics.

Analysing user preferences in order to improve applications and user experiences on the Web is in fact one of the well-researched areas in computer science. A prominent example is the study of relevance feedback for enhancing Web search and information retrieval (Joachims, 2002; Joachims et al., 2005; Vassilvitskii and Brill, 2006), in which the preferences of users—the items preferred by the users—after they have submitted a search query are collected either explicitly or implicitly for improving the search results. It suggests that very often even ranking algorithms such as PageRank (Brin and Page, 1998) and keyword-based information retrieval techniques are not good enough at producing a list of documents that can satisfy the information needs of the users. The reason is clear: these methods usually only focus on the characteristics of the documents themselves, and do not consider the perspectives of the users. A recent study by Agrahri et al. (2008) shows that 'people's shared preferences do not always agree with Google's result order'. Hence, for Web applications to be useful to the users, information should be processed and presented in a way that takes their perspectives into ac-

count. In this respect, the Social Web and especially collaborative tagging systems offer great opportunities for harnessing user preferences to improve Web applications, including organising, searching and sharing of online resources. A better understanding of the mechanisms of collective semantics is therefore essential in the process of transferring knowledge of user preferences to implementations that facilitate various activities on the Web.

Collaborative tagging can be considered as an excellent component of the Social Web for studying collective semantics. It involves simple interactions, but captures all the important entities of the Social Web. Collaborative tagging systems are also sources of rich information about user preferences. Unlike records in a Web log or a search engine query log, a folksonomy provides more reliable evidence of the positive associations between users and Web documents. A user is more likely to be interested in a document if he/she has tagged it in a folksonomy than if he/she has only browsed it or clicked on it after submitting a query. Collaborative tagging is also very popular nowadays, and as a result it offers a large volume of data that covers a wide range of domains for analysis and experimentation. In this thesis, we refer mainly to collaborative tagging data obtained from the popular social bookmarking system Delicious (see Chapter 3 for a more detailed description). In the following section, we describe several research questions in relation to collective semantics that we will investigate in this thesis.

1.4 Research Questions and Hypotheses

This thesis centres around the notion of collective semantics on the Social Web with emphasis on collaborative tagging. By conducting the research work described in this thesis, we aim at answering the following research questions:

- How can we understand and extract the collective semantics of the entities found in a collaborative tagging system?
- How can the collective semantics discovered in a collaborative tagging system benefit the users?

These two questions are very general, but are central to our thesis. Firstly, we want to know the methods we can rely on to extract the collective semantics of the entities in a folksonomy. This involves how we model the data collected, what kinds of algorithms do we employ to analyse the data, and how do we evaluate our

methods. Secondly, we want to know whether the collective semantics discovered is useful from the perspective of the users. For example, can we use the findings to improve user experience in collaborative tagging or even in other applications on the Web? Can we use this better understanding of the entities in a folksonomy to improve retrieval of and navigation among documents? Ultimately, answering these questions allow us to gain a better understanding of the Social Web and to improve Social Web applications to facilitate user interactions on the Web.

As collaborative tagging involves three different types of entities, namely tags, users and documents, we attempt to answer the two general research questions by conducting research work on three corresponding dimensions.

Firstly, we will study the semantics of tags. Tags are important entities in a folksonomy because they describe what a document is about and what a user is interested in. However, the tags themselves are ambiguous in the sense that there are no rules as to what a tag should be used to represent and what the relations between different tags are, due to the flexibility of collaborative tagging. Without such restrictions, a tag can be used to represent anything the users intend without adhering to the conventional meaning(s) of the tag. Ambiguity of tags can be a problem if we rely on them to retrieve relevant documents or as a means of navigation through documents of related topics. We believe relying on external resources such as dictionaries may not be very helpful because the vocabulary in a folksonomy is very dynamic, such that new words are constantly introduced and existing words can be used with new meanings. Instead, we believe that tag contextualisation—the act of identifying the different contexts in which a tag is used—can be done by examining the complex network structure resulted from the collective user behaviour in a collaborative tagging system. We also believe that by contextualising tags we can improve organisation and retrieval of resources on the Web. In particular, we test the following hypothesis regarding meanings of tags in collaborative tagging:

> **Hypothesis 1 (tag meaning)**: When modelling a folksonomy, networks that explicitly take the users' collective behaviour into account are better in capturing meaningful associations between different entities in a collaborative tagging system, and clustering analysis of these networks produces more accurate results regarding the meanings of tags intended by the users.

The second type of entities in a folksonomy we will investigate is users. Users are different from the other two types of entities because they are the ones who determine what associations between tags and documents are established. They also actively introduce new tags and documents into the folksonomy. Talking about 'semantics' of users may sound strange. However, as we have mentioned earlier, we use the phrase 'collective semantics' to refer not only to the meanings of words but also the qualities or characteristics of the entities we are studying. Hence, we can also study the collective semantics of the users. As users are the ones who introduce new resources into a folksonomy, a quality of users that we are interested in is their trustworthiness. Here, an important question would be to what extent we can rely on the input of a user. Note that in many tagging systems users are only characterised by their usernames and the tags and documents they have contributed. In other words, there is no information about the credibility or the expertise of a user. However, we believe that the trustworthiness/credibility/expertise of the users can be understood by studying the implicit interactions between themselves. By developing a method to rank users based on their implicit interactions, we also believe that we can reduce the negative impact of malicious users on the system. In summary, we want to test the following hypothesis regarding users in collaborative tagging:

> **Hypothesis 2 (user expertise)**: The trustworthiness or expertise of the users in a collaborative tagging system can be derived from analysis of the implicit interactions among the users themselves in their tagging activities.

Finally, we will study documents found in a folksonomy. Relations between documents on the Web are generally defined by the hyperlinks between them. Documents connected by hyperlinks can be considered as related to each other in some ways. On the Web, a user following hyperlinks from one document to another will find relevant or complementary information that cannot be found within a single document. In other words, the relations between documents are particularly important when their 'semantics' is considered. Collaborative tagging allows users to establish their own collections of documents described by their own tags. The tagging activities of users therefore generate some implicit relations between the documents. For example, two documents that are both interesting to a large number of user or have both been assigned similar sets of tags can be considered as highly related to each other, no matter whether they are connected by hyperlinks or not. The important questions here are how we can identify these implicit

relations between the documents, and how these relations are different from the relations defined by existing hyperlinks. We believe that such implicit relations should act as better links between documents to direct users to relevant information, because they are generated from the perspectives of the users, as opposed to the hyperlinks created by the authors of the documents. In particular, the following hypothesis will be tested:

> **Hypothesis 3 (relations between documents)**: The implicit relations between documents generated by the collective tagging activities of Web users represent better recommendation links than existing hyperlinks created by the authors of these documents.

While we discuss these three dimensions separately, they are closely related to each other and are highly relevant to our central theme of this thesis. We aim at studying how implicit relations between entities resulted from the collective behaviours of the users in a collaborative tagging system can be analysed to reveal the semantics of the entities. All these studies contribute to answering our research questions presented in the beginning of this section: how can we understood and extract the collective semantics of entities in a folksonomy and how does this information benefit users of collaborative tagging systems.

1.5 Contributions

This thesis presents empirical studies of collaborative tagging and folksonomies. In particular, it investigates the notion of collective semantics on the Social Web with reference to real-world data collected from the exemplar collaborative tagging systems Delicious. The contributions of this thesis include the followings:

- We put forward the notion of collective semantics in the context of the Social Web, which gives us a framework for understanding the characteristics of the three types of entities on the Web.
- We present a comprehensive review of previous research works and studies on collaborative tagging and folksonomies. We also discuss the relations between traditional subject indexing and collaborative tagging. (Chapter 2)
- We study thoroughly the problem of tag contextualisation by comparing the effectiveness of different network representations of a folksonomy at the level

of individual tags. We find that clustering analysis of networks that explicitly take users' collective behaviour into account return results that are more accurate in revealing the meanings of tags intended by the users, supporting our hypothesis (Hypothesis 1) regarding tag meanings. We also experiment with the idea of using the outcomes of tag contextualisation to assist Web search result classification in order to improve performance of information retrieval on the Web. We propose a k-nearest-neighbour classification method for the purpose, and show that it gives promising results. (Chapter 4)

- We discuss the notion of implicit endorsement in collaborative tagging, and propose an algorithm, SPEAR, that implements the idea for ranking users according to their expertise. By analysing the results of simulations based on models of different types of users and qualitatively examining results returned by SPEAR and other algorithms, we show that certain qualities (trustworthiness/expertise) of the users who participate in a collaborative tagging system can be revealed by analysing the implicit interactions between themselves, providing strong support to our hypothesis regarding user expertise (Hypothesis 2). (Chapter 5)

- We propose two different approaches for identifying implicit relations between documents in a folksonomy. We compare these implicit relations with existing hyperlinks and find out that the former relations represent better links that allow users to discover other relevant and useful documents, which supports our hypothesis regarding relations between documents in collaborative tagging (Hypothesis 3). These user-induced links provide a different perspective to the relations between documents compared to existing hyperlinks. We also experiment with predicting the tags of a document by using implicit links generated from a folksonomy. Our experiments show that they are very useful and lead to predictions of high accuracy. This suggests that relations based on user preferences are also useful in classification tasks. (Chapter 6)

In addition, earlier versions of several different parts of this thesis have been published and presented in international conferences in the past few years. These include ISWC+ASWC 2007 (Au Yeung et al., 2007a,c), WI+IAT 2007 (Au Yeung et al., 2007b), ECIR 2008 (Au Yeung et al., 2008e), WI+IAT 2008 (Au Yeung et al., 2008a,b,c), WebSci 2009 (Au Yeung et al., 2009e), ACM HYPERTEXT 2009 (Au Yeung et al., 2009a), ACM SIGIR 2009 (Noll et al., 2009) and ACM CIKM 2009 (Au Yeung et al., 2009c). A related study on the diversity of user

interests in collaborative tagging systems has been published and presented in a workshop in WWW 2008 (Au Yeung et al., 2008d), and subsequently extended to a book chapter (Au Yeung et al., 2009b).

1.6 Structure of the Thesis

The structure of this thesis is as follows.

Following this introductory chapter, Chapter 2 will present a literature review, which will include a detailed description of the characteristics of existing collaborative tagging systems and a thorough review of the state-of-the-art research projects and studies of collaborative tagging and folksonomies. Chapter 3 will give a description of the subject of study of this thesis, the popular collaborative tagging system Delicious. It will also describe our process of data collection and provide an overview of the collected data sets. Chapter 4 will study the semantics of tags found in our data sets, discuss the nature of tags that have multiple meanings, investigate how the social meanings of tags can be discovered by cluster analysis of different network representation of a folksonomy, and finally present an experiment on how the multiple meanings of tags discovered can be applied to classify Web search results. Chapter 5 will explore the notions of implicit endorsement and expertise in the context of collaborative tagging. It will describe our proposed algorithm for ranking users in a folksonomy by their expertise, based on an implementation of the notion of implicit endorsement. It will also present our experiments on the effectiveness of the proposed algorithm and our analysis of the experimental results. Chapter 6 will investigate how implicit relations between documents resulted from collaborative tagging activities can be identified through analysis of document similarity and user preferences. We will compare these implicit relations, what we call user-induced links, with existing hyperlinks. The chapter also presents our proposal of using user-induced links to help predict the tags of a document. Finally, Chapter 7 will give concluding remarks for this thesis. We will highlight important findings and discuss the implications and significance of the research work described in this thesis. We will also at the end outline possible research directions.

Chapter 2

A Review of Collaborative Tagging

Collaborative tagging represents a new means of organising and sharing resources on the Web. Its popularity among Web users results in a large amount of data available for analysis, leading to research works that investigate various characteristics of tagging systems and study different methods of extracting useful information from the data. In this chapter we present a thorough review of collaborative tagging, including the idea of using keywords to describe resources and the various characteristics of collaborative tagging systems. In addition, we summarise the results of recent analytical studies and research works on collaborative tagging systems. It should be noted that collaborative tagging has been a popular social phenomenon (at least on the Web), and therefore it has attracted the attention of researchers from a wide range of domains, including for example library science, media studies, physics and, of course, computer science. Although we are studying how the social and collective behaviour of users in collaborative tagging systems, we approach the issue from a scientific and mathematical perspective. Therefore, we mainly focus on the literature of computer science here, and refer to studies in other research domains where necessary and appropriate.

2.1 Subject Indexing of Documents

While collaborative tagging has only become popular among Web users in recent years, the basic idea underlying the notion of using tags to describe documents has actually been around for quite a long time, and has been studied under the

name of subject indexing in various fields. Subject indexing (Lancaster, 2003) refers to the task of constructing a representation of a resource—a document, a photo, a video tape, etc.—in order to facilitate its retrieval at a later time. While there can be different motivations behind assigning tags to documents on the Web, collaborative tagging can generally be regarded as a form of manual indexing (Voss, 2007).

Subject indexing is never a trivial topic because there are no correct answers as to how a particular item should be indexed. According to Lancaster (2003), the task of subject indexing involves two major steps, namely conceptual analysis and translation. In the process of conceptual analysis, the indexer (usually a human being possessing relevant knowledge) figures out the topics addressed by the item. In the process of translation, the indexer decides on the set of index terms, which are also known as keywords, based on the result of conceptual analysis. Effective subject indexing therefore usually involves a correct decision on what an item is about and appropriate choices of index terms. While traditional subject indexing usually involves a small group of indexers assigning index terms to items based on a predefined vocabulary or taxonomy, some suggest that taking the interests of the users into consideration is also necessary. After all, what is important is whether users will be able to retrieve the items easily in the future. For example, Fidel (1994) and Layne (2002) discuss the differences between 'document-oriented' indexing and 'user-oriented' indexing. They suggest that there are needs for different indexing with different terminology for different audiences. This idea is actually quite similar to that of collaborative tagging in which users are usually allowed to maintain their own tags for the items.

In the context of the Web, Yahoo! represents one of the earliest attempts to index and categorise documents on the Web. The Yahoo! Directory involves predefined categories and Web pages are classified to these categories manually by staff members of the company.[1] The Open Directory Project is a similar example except that it is constructed by a huge community of volunteer editors from different parts of the world.[2] However, as the size of the Web continues to grow at a rapid rate, this kind of manual indexing of Web documents is obviously not efficient. Google is no doubt the most prominent search engine that tackles the huge volume of information on the Web by indexing documents on the Web in an automatic fashion, extracting keywords from the documents themselves as index terms and enhancing the retrieval process by adopting the PageRank link analysis algorithm

[1] Yahoo! Directory: `http://dir.yahoo.com/`

[2] ODP- Open Directory Project: `http://www.dmoz.org/`

(Brin and Page, 1998). Interestingly, collaborative tagging systems that have become so popular recently seem to go back to the early days of subject indexing as they promote manual indexing of Web documents (Voss, 2007).

In fact, there are actually some proposals in the literature that suggest user-oriented and distributed indexing methods similar to that of collaborative tagging. For example, Brown et al. (1996) discuss the idea of 'democratic indexing' in the context of image indexing. They suggest that in addition to index terms that have been assigned to an image, users should be allowed to add their own terms where necessary and appropriate to facilitate organisation and retrieval. Along a similar line of thought, Besser (1997) proposes developing systems for user-assigned terminology, systems that allow users to assign terms or keywords to individual images, as a solution to the scarcity of metadata of images. He also suggests that users of such systems can limit their searches to terms assigned by people who they trust, either because they possess more relevant knowledge or their contribution are considered more reliable. This is actually highly relevant to nowadays collaborative tagging systems as gaming and spamming activities are found to be very common in these systems (Wetzker et al., 2008). In addition, Villarroel et al. (2002) propose a system in which weights of the index terms of a document are revised according to which parts in the text are highlighted by the users, thus implicitly allowing users to determine which terms are more important to the document. A system that is very similar to nowadays social bookmarking system is proposed by Keller et al. (1997). The system is a Web-based bookmarking service implemented by means of a proxy server. A user can submit a bookmark to the service and assign it to some categories, which can be created without any restriction to a predefined vocabulary or hierarchical structure. Every users from the same group can collaboratively maintain the set of bookmarks within the system. It also incorporate user feedbacks that are used to rank bookmarks with respect to the categories.

Obviously, the basic notion underlying collaborative tagging has been around for quite a long time. It is interesting that such idea of user-oriented indexing was not realised and popularised until nearly ten years after Brown et al. (1996) first proposed this approach. This, to a large extent, can be attributed to several reasons. Firstly, many previous proposals mentioned above mainly focus on some small, closed and stand-alone systems, whereas the Web provides a huge platform to facilitate the widespread participation of users. The ability to solicit input from users from all over the world also creates a richer and more diverse collection of resources. Secondly, nowadays collaborative tagging systems place more emphasis

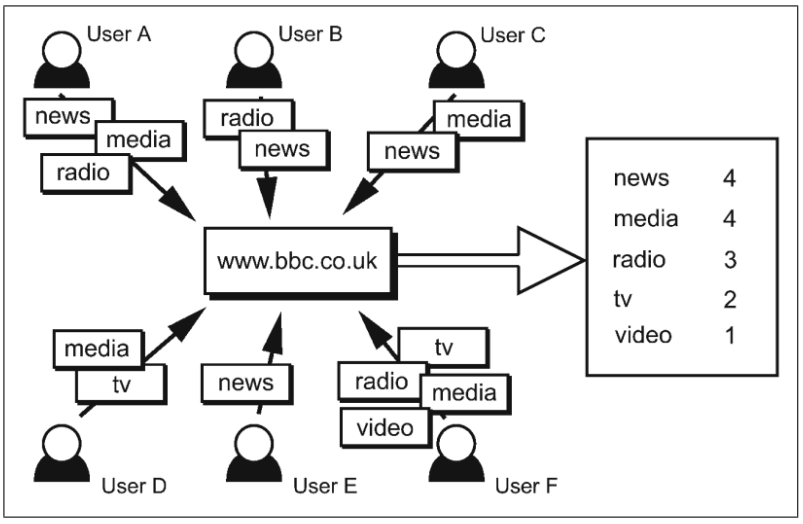


FIGURE 2.1: An example of how a typical collaborative tagging system works.

on the social aspect of indexing, allowing users to instantly see how their annotations compare with those of the others. Users also benefit immediately from the annotations of the others as any tags assigned to a document can be used for retrieval and navigation. Thirdly, in recent year, a number of new design principles promoted by Web 2.0 produce user interfaces that are more interactive and easy to use. While it is true that the technical standards involved in Web 2.0 applications are nothing new, improved design coupled by the prevalence of Web browser plugins help to a very large extent popularise collaborative tagging on the Web.

2.2 Collaborative Tagging Systems

Collaborative tagging systems are Web-based systems which allow users to assign tags to resources available on the Web for the purpose of organisation or sharing with other users. While subtle differences exist across different systems, tags are in general keywords or phrases which are created on-the-fly by users with few restrictions. The act of assigning a tag to an object is referred to as ***tagging***. Figure 2.1 presents an example of how a typical collaborative tagging system works: users assign tags in the form of descriptive keywords to a document they are interested in (in this case the homepage of the British Broadcasting Corporation).

According to Hammond et al. (2005), collaborative tagging systems have started to thrive and grow in number since late 2003 and early 2004. Prominent examples include Delicious, a social bookmarking system which allows Web users to store

Name	**URL**	**Resource Type**
Delicious	http://delicious.com/	bookmarks
Flickr	http://www.flickr.com/	photos, images and videos
BibSonomy	http://www.bibsonomy.org/	bookmarks, academic references
CiteULike	http://www.citeulike.org/	academic references
LibraryThing	http://www.librarything.com/	books

TABLE 2.1: Some popular collaborative tagging systems on the Web.

their bookmarks (shortcut to their favourite Web pages) in its system and describe them with tags; Flickr, which provides hosting services of digital images and allows users to describe them with tags; and LibraryThing, which is similar to Delicious except that the objects being tagged are books instead of Web pages. Table 2.1 provides a list of some of the most popular collaborative tagging systems. It should be noted that while these systems first appeared in English-speaking communities, there are also systems that are designed for particular language groups, such as the Japanese site Hatena[3] and the Korean site Margarin[4].

As we have mentioned in the previous section, collaborative tagging is a special form of manual subject indexing, as it involves human users assigning keywords to resources for the purpose of organisation and future retrieval. However, collaborative tagging has some significant differences when compared with traditional subject indexing. These include the following three major aspects.

- **No controlled vocabulary.** Traditional subject indexing usually involves selecting terms from a predefined vocabulary or taxonomy. For example, the Library of Congress Classification (LCC) system is used to assign subject headings to books in libraries.[5] The Yahoo! Directory represents a similar effort in the context of the Web. However, users do not choose from any controlled vocabulary when assigning tags to resources. Instead, they usually have the freedom to come up with any keywords or phrases for the purpose, conforming only to a minimal set of rules (e.g. no space within a tag) depending on the design of the collaborative tagging system.

- **Indexing by users.** In traditional subject indexing, index terms are usually chosen and assigned by the creator of a resource or by experts with relevant domain knowledge. However, in collaborative tagging systems, it is the users who assign tags to resources they are interested in. In other words, tags

[3]Hatena: `http://www.hatena.ne.jp/`

[4]Margarin: `http://mar.gar.in/`

[5]The Library of Congress Classification: `http://www.loc.gov/aba/cataloging/classification/`

represent annotations of a resource from the perspective of the readers or the consumers of the resource, which can be very different from the index terms assigned by the authors of the resource or an expert of the classification system.

- **Multiple sets of index terms.** As its name suggests, collaborative tagging involves assigning tags to resources in a collaborative manner, i.e. a resource can be assigned tags by more than one user. Most collaborative tagging systems (e.g. Delicious) allow every user to maintain his/her own set of tags with respect to a certain resource. A user therefore does not need to negotiate with other users to arrive at a consensus before he/she assigns his/her tags to a resource (although implicit consensuses can usually be found in these systems as we will discuss later in this chapter).

It should also be noted that collaborative tagging systems usually publish their data on the Web such that the tagging history of a particular item or user can be traced.[6] This is unlike some online bookmarking services which allow users to store their bookmarks in private accounts for the purpose of being able to retrieve the bookmarks from different computers. This characteristic produces a sense of social collaboration in tagging. While users may still use tags primarily for the purpose of organising their favourite resources and facilitating future retrieval, the collaborative nature of tagging on the Web also results in continuously evolving classification scheme. Considering the example in Figure 2.1 again, different users assign different tags to the URL `http://www.bbc.co.uk/`. When the data is aggregated by the collaborative tagging system, it can be observed that the tags `news` and `media` are used by the greatest number of users. These common tags can then be considered as the users' opinion on how this Web page should be classified. Such classification observed in collaborative tagging systems are now commonly referred to as a ***folksonomy***, which we will discuss in more detail later in this chapter.

2.2.1 Design Issues

While collaborative tagging systems are all based on the simple idea of allowing users to describe Web resources with freely-chosen tags, different collaborative tagging systems are designed by adopting different design principles which render

[6]The amount of data published varies from system to system. For example, Delicious only shows up to 2,000 records for a particular tag after it has launched a new interface in July 2008.

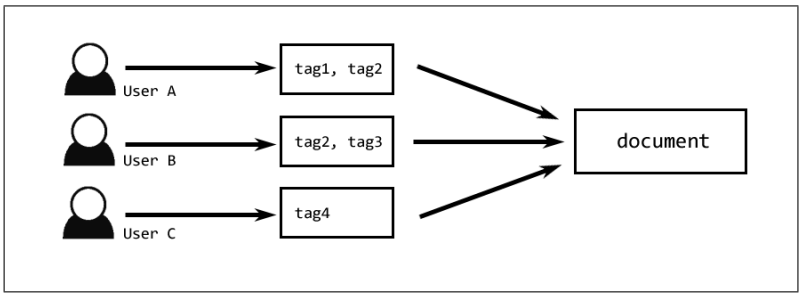
Dimension	**Description**
Tagging rights	Who can tag the resources? (self, permission-based, free-for-all)
Tagging support	How are the users support? (e.g. suggesting popular tags)
Aggregation model	How are tags aggregated? (recording the frequency of a tag or not)
Object type	What are the resources? (URLs, images, books, etc.)
Source of material	Where do the resources come from? (owned by users or publicly available)
Resource connectivity	Can resources in the system be grouped or linked?
Social connectivity	Can users interact directly with other users?

TABLE 2.2: Dimensions of tagging system design and their descriptions (Marlow et al., 2006).

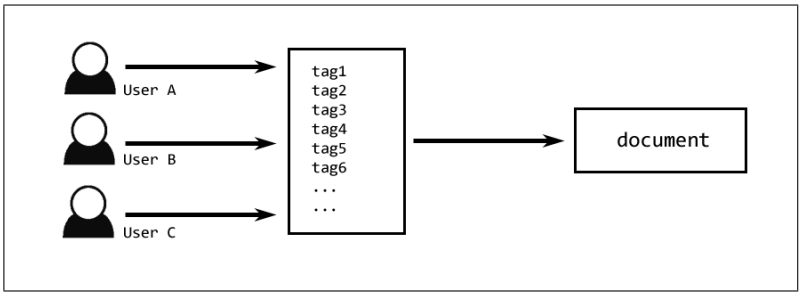
different functionalities and characteristics. Table 2.2 shows a list of dimensions of tagging system design presented by Marlow et al. (2006). Sen et al. (2006) describe an experiment on vocabulary formation in the MovieLens system, the authors show that different design choices affect the nature/types of tags used, their distributions and the convergence within a group.

One obvious example of the different dimensions of system design is the type of resources being tagged. The types of resources being tagged in the system generally affect the overall design of the system. In addition, it is found that the types of tags (see Section 2.3.2.4) that are used heavily by the users depend on the types of resources being tagged in the system (Bischoff et al., 2008). Another dimension on which systems differ is the ownership of the resources being tagged. For example, in Delicious the resources are URLs, each of which represents a document on the Web. Hence, although the documents are in general publicly available, the users usually do not own the documents. This also applies to LibraryThings in which the items being tagged are books. Flickr, on the other hand, is specialised in tagging of digital images, and these images mainly originated from the users themselves.

Another obvious difference between different collaborative tagging systems can be found in their models of tagging. The choice of a particular model determines who has the right to assign tags to resources, how tags are aggregated and the type of folksonomy which results from the collective tagging activity (Marlow et al., 2006; Sen et al., 2006). For example, any user is allowed to assign tags to any resources posted to the system in Delicious. Every user of Delicious maintains a unique set of tags for their favourite bookmarks. In Flickr, however, only the owners of the images or their friends have the right to assign tags. Moreover, the



(a) A broad folksonomy refers to the model in which users maintain their own sets of tags for each document.



(b) In a narrow folksonomy, only one set of tags is maintained for each document in the system.

FIGURE 2.2: Two types of folksonomies.

system only maintains a single set of tags for each item. This kind of difference, as we will see, leads to very different tagging behaviour and produces different types of folksonomies. For systems like Delicious in which every user is allowed to maintain their own set of tags, a ***broad folksonomy*** is produced (see Figure 2.2(a)); for systems like Flickr in which only one set of tags is maintained for each item, a ***narrow folksonomy*** is produced (see Figure 2.2(b)) (Vander Wal, 2005). In other words, a narrow folksonomy contains no record of how many times a particular tag has been used on a document.

There are also other dimensions on which collaborative tagging systems differ, such as interface design and other functionalities of the systems. Systems may support users when a new item is added by suggesting popular tags which have already been assigned to the item by previous users. Some system such as BibSonomy allow user to define simple hierarchical relations between tags. Subtle distinctions such as the choice of delimiter used when assigning tags and whether spaces are allowed to appear in tags can also be found across different collaborative tagging systems. In addition, some systems feature functionalities which foster the development of social networks. For example, Delicious allows users to establish social networks by subscribing to the RSS feeds of other users and by keeping track of who are subscribing to one's own feeds. Flickr also allows users to label other users as friends or family members to grant permissions to further tag or comment on

their photos and videos.

2.3 Folksonomies

As more and more users contribute their tags to a collaborative tagging system, a form of classification scheme takes shape. Such a scheme results from the collective efforts of the participating users, reflecting their viewpoints on how the shared resources on the Web should be described using different tags. This product of collaborative tagging is now commonly referred to as folksonomy, a term first proposed by Vander Wal (2005) as a combination of the terms 'folk' and 'taxonomy'. Folksonomy has been given other names throughout the development of collaborative tagging systems, such as *folk classification*, *distributed classification*, *social classification*, *faceted hierarchy* and *ethnoclassification* (Hammond et al., 2005). Since the term folksonomy has been mostly used among both Web users and researchers of the topic, it will be used in this thesis to refer to the aforementioned scheme generated in collaborative tagging systems.

Folksonomies consist of tags that are assigned by users following no specific guidelines or rules and are therefore usually considered to be unorganised and chaotic. However, because users are supposed to assign tags that help them to organise and retrieve their favourite resources (Golder and Huberman, 2006), the relations between the three types of elements—users, tags and documents—in a folksonomy are therefore not arbitrary. By analysing these interrelations it is possible to discover much implicit information about these elements. In fact, it is shown that folksonomy can be used to construct light-weight ontologies from a bottom-up approach (e.g. Mika 2007; Zhou et al. 2007b), to enhance recommendation systems by treating tags as annotations of documents (e.g. Niwa et al. 2006; Shiratsuchi et al. 2006), or to generate profiles which represent the interests of the Web users (e.g Michlmayr and Cayzer 2007). In this section we review different analyses of folksonomies and different applications of the data found in a folksonomy.

2.3.1 Formal Models of Folksonomies

To discuss folksonomies in mathematical terms, a formal model of folksonomies is usually required. In general, a folksonomy is consider to consist of at least the following three sets of entities (Marlow et al., 2006; Mika, 2007; Wu et al., 2006).

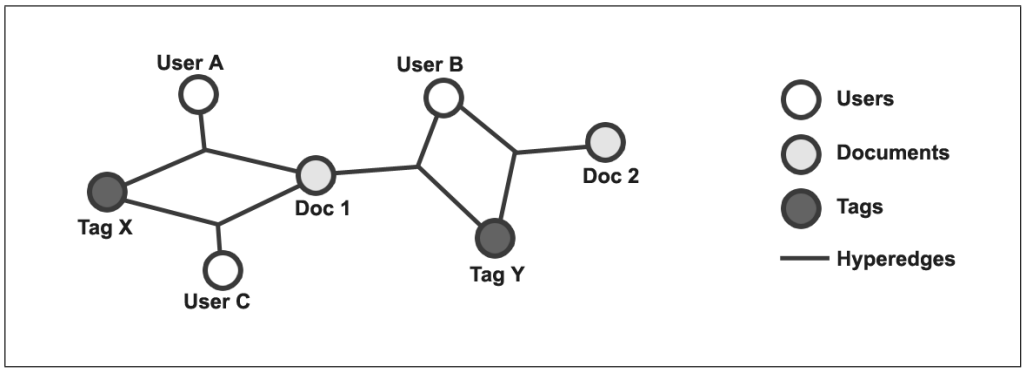


FIGURE 2.3: The hypergraph of a simple folksonomy. There are three disjoint sets of entities, namely users, documents and tags. Each hyperedge connects a member from each of these sets. For example, in the above folksonomy, User A has used Tag X to describe Document 1.

- **Users** assign tags to Web resources in social tagging systems. They are also referred to as actors, as in social network analysis, or agents as in multi-agent research.

- **Tags** are keywords chosen by users to describe and categorise Web resources. Depending on systems, tags can be a single word, a phase or a combination of symbols and alphabets. Tags are referred to as concepts in some works which focus on extracting lightweight ontologies from folksonomies.

- **Resources** refer to the items which are being tagged by users in collaborative tagging systems. Depending on the system, there are a wide range of resources being tagged, such as Web pages (bookmarks) in Delicious, digital images in Flickr, and video clips in YouTube. Resources are also referred to as instances, objects or documents, depending on the context.

Mika (2007) represents a social tagging system as a tripartite hypergraph (see Figure 2.3), in which the set of vertices can be partitioned into three disjoint sets A, C and I, corresponding to the set of actors, the set of concepts and the set of objects being tagged. A folksonomy is then defined by a set of annotations $T \subseteq A \times C \times I$, an element of which is a triple representing the fact that an actor has assigned a concept to an object.

Hotho et al. (2006a) define a folksonomy as a tuple $\mathcal{F} := (U, T, R, Y, \prec)$. The finite sets U, T and R correspond to the set of users, tags and resources respectively. Y refers to the tag assignments, which are ternary relations between the above three sets: $Y \subseteq U \times T \times R$. Finally, $\prec$ is a user-specific relation which defines the sub/superordinate relations between tags. This relation can be found in some

collaborative tagging systems such as BibSonomy. By dropping the relation $\prec$, the folksonomy can be reduced to a tripartite hypergraph, which is equivalent to the model used by Mika.

In fact, except a few studies which focus on analysing changes and trends found in a folksonomy across time (Hotho et al., 2006b), most studies of folksonomy in the literature refer to the basic model involving the three basic entities. In this thesis, we focus on the interrelations between the three basic entities in Delicious (more on Delicious in Chapter 3), other information such as the time at which a tag is assigned or the subsumption relations between tags will not be considered. Hence, we adopt the following simple model of folksonomy in this thesis.

Definition 2.1. A folksonomy $\mathcal{F}$ is defined as a tuple $\mathcal{F} = (U, T, D, R)$, where U is a set of users, T is a set of tags, D is a set of documents, and $R \subseteq U \times T \times D$ is a set of annotations.

2.3.2 Characteristics of Folksonomies

Folksonomies are primarily produced by the collective actions taken by users participating in collaborative tagging. As a result, the characteristics of folksonomies depend on the dynamics of the behaviour of the participating users. In addition, there has been wide discussion about the advantages and disadvantages of using folksonomies as a means of organising and annotating resources on the Web. In this section we discuss the strengths and weaknesses of folksonomies and go on to examine some well-researched characteristics of folksonomies.

2.3.2.1 Strengths

Quite a number of studies review the characteristics of collaborative tagging systems and folksonomies (Quintarelli, 2005; Furnas et al., 2006; Wu et al., 2006), in which reasons of why these systems are so popular and are widely accepted by the general users are discussed. The following features of collaborative tagging and folksonomies are generally attributed to its success and popularity.

- **Low cognitive cost and entry barriers**. Contrasting to cases in which semantic annotation is done by referring to some ontologies, the simplicity of tagging allow any Web users to classify their favourite Web resources using

freely-chosen keywords which are not confined to any predefined vocabularies. Users do not need to possess any prior knowledge of a particular domain or the hierarchical structures of certain taxonomies in order to participate in tagging activities. Besides, users feel much more comfortable in using their own words to describe their favourite resources.

- **Better retrieval of resources**. As mentioned in the previous point, users can use terms that they feel most comfortable and familiar with to label resources. In this way, when the user wants to retrieve the stored information, he/she can obtain the information more quickly by using his/her own words to describe these relevant concepts, instead of being forced to retrieve a set of resources by using pre-defined categories.

- **Immediate feedback and communication**. When a user assigns tags to a Web resource, many existing tagging systems will give information on what kind of tags other users have already used to describe this resource. Through this mechanism, users can modify their own choice of tags by following the general practice of other users, or they can choose to insist on their choice so as to influence the other users. In other words, users usually receive immediate feedback from the system on their actions. In addition, collaborative tagging represents a kind of implicit communication between the users of a system with the common goal of coming up with a set of suitable tags for a particular resource.

- **Quick adaptation to changes in vocabulary**. Traditional classification schemes such as taxonomies or ontologies are usually slow in responding to changes in the use of language and the emergence of new words. In contrast, users are free to create any words to tag resources on the Web in collaborative tagging systems. Terms like 'ajax', 'web2.0', 'ontologies' and 'socialnetwork' can be readily used without the need to modify any predefined schemes. In this way, new terms find their way into the system very quickly. This also implies that collaborative tagging systems are able to accommodate new information very quickly.

- **Individual needs and formation of organisation**. Tagging systems provide Web users a convenient means to organise their favourite Web resources by simply assigning tags to them. The advantage of being able to classify as well as retrieve these resources based entirely on the tags chosen by themselves contributes to boosting the incentive to use these systems. Besides, as the folksonomy evolves, users are able to discover other people who are

also interested in similar items, and eventually they can share more of their favourite resources as well as their choices of words on tagging.

2.3.2.2 Weaknesses

On the other hand, limitations and problems of existing collaborative tagging systems and folksonomies have also been identified (Niwa et al., 2006; Wu et al., 2006). The following issues are commonly regarded as problems or weaknesses of collaborative tagging when compared with more structured or formal approaches of information organisation.

- **Ambiguity of tags**. Tags in collaborative tagging systems are created freely by users at will. As a result, there is no way to make sure that a tag always corresponds to a single and well-defined concept. In other words, a tag can be ambiguous. For example, items assigned the tag `sf` in Delicious include pages about the city of San Francisco as well as pages about science fictions. This clearly suggests that the tag `sf` is used by users to refer to two distinctive concepts: a city in the United States and a genre of books. Many other examples can easily be discovered. Such ambiguity of tags is a problem when tags are used to retrieve resources stored in a collaborative tagging system (Golder and Huberman, 2006), since the systems would not be able to differentiate the different concepts that are being referred to by the same tag.

- **The use of multiple words and spaces**. Most existing systems allow users to assign tags to Web resources by typing in a list of tags separated by spaces. In other words, spaces are treated as delimiters of tags in the input field. In many situations, users would choose a phase with multiple words as a tag, without knowing that the system interprets the phase as separate words for different tags. This results in a lot of tags that cannot be understood independently. In order to use a phase for a tag, users therefore need to type the phase without spaces, or use capital first letters. Again, this causes problems when one attempts to retrieve resources by tags.

- **The problem of synonyms**. In addition to the fact that a tag is used to refer to different concepts by different users, different tags may also be used to refer to the same concept in a tagging system. For example, `mac`, `macintosh`, and `apple` are all used to describe Web resources related to Apple Macintosh

computers in Delicious. In addition, as collaborative tagging systems usually recognise and store tags in the same form as they are created, tags of the same word stem are still regarded as different tags. Examples are singular versus plural nouns (e.g. `game` vs. `games`), words of different parts of speech (e.g. `blog` vs. `blogging`), and spelling differences (e.g. `center` vs. `centre`). Synonyms may cause problems in retrieval of resources by tags as one may need to specify all synonyms in order to obtain all the relevant resources.

- **Lack of semantics**. Although tagging systems allow Web users to associate tags with Web resources, the types of these associations cannot be specified. In general, when a tag t is associated with a document d, the relation only means that 'd is about t', while there are actually much more possibilities. For example, when a Web page is assigned the tag `dickens`, it may be the case that the page contains a biography of Charles Dickens, or it may be the case that the page contains a novel written by Charles Dickens, not to mention the possibility that the page is about another person with the same name. In this respect, a tag provides limited information about the content of a resource.

Given the contrasts between folksonomies and controlled vocabularies, as well as the advantages and limitations of collaborative tagging, there are both advocates and critics of folksonomies. In a controversial article published on the Web, Shirky (2005) suggests that folksonomies should be favoured to solve many existing problems of classification that use predefined taxonomies or ontologies. This is because folksonomies provide a bottom-up approach for making sense of the resources available on the Web from the users' perspective. He/she also suggests that synonyms may not actually be a problem in collaborative tagging because users always use a particular tag for some reasons, and different tags always contain some useful information. On the other end of the spectrum, Peterson (2006) points out from a philosophical perspective that folksonomies suffer from limitations similar to those of relativism. While folksonomies allow every user to have a say of how a particular object should be classified, folksonomies are prone to inaccuracy, inconsistency and contradictions. However, Crawford (2006) and Guy and Tonkin (2006) suggest that traditional classification systems and folksonomies should not be seen as a dichotomy, and they should be used to complement each other. We will discuss in Section 2.3.7 studies that attempt to generate ontologies from folksonomies or to enrich tags in folksonomies with semantics in existing ontologies.

In fact, the advantages and limitations of collaborative tagging can only be judged fairly when they are discussed with respect to some particular application. Clearly, the Web involves a very large number of users interested in different domains and a huge volume of information of different topics. In such a diverse setting it would be very difficult, if not impossible, to ask all users to refer to a common vocabulary when describing and organising their favourite Web resources, not to mention the difficulty of coming up with this vocabulary. Collaborative tagging thus represents a flexible and scalable method for general Web users to annotate Web resources for their own purposes as well as for sharing with the others. While users may still use tags in a subjective and idiosyncratic way, it can be observed from existing systems that consensus does emerge from the collaborative tagging behaviour of the users (see Section 2.3.2.3). On the other hand, well-defined taxonomies and ontologies would similarly find their usage in facilitating organisation and retrieval of information in other application scenarios, which may involve a more refined domain of knowledge in which consensus can be reached more easily or is very much needed.

Given the popularity and some of the advantages of folksonomies, we already see quite a number of cases in which collaborative tagging is used alongside traditional classification approaches to provide better information management. For example, tags contributed by users of the library at the University of Pennsylvania are used to help organising books in the library along with the existing subject indexing system.[7] It has also been reported that social tagging is used on museum collections for classification by combining the perspectives of the visitors and those of the museum curators (Trant, 2006; Trant and Wyman, 2006). In addition, IBM is reported to have implemented a social bookmarking tool called Dogear for use within an enterprise (Millen et al., 2006). These applications suggest that the combination of collaborative tagging and traditional classification approaches is feasible and beneficial in many different settings.

2.3.2.3 Usage Patterns

In one of the earliest studies of collaborative tagging, Golder and Huberman (2006) report a major characteristic of the folksonomy in Delicious which is least expected. The design of Delicious allows users to use any words they like as tags to describe documents on the Web. Since users are not required to follow any specific guidelines, it is natural to suggest that a chaotic pattern of tags would be observed.

[7] PennTags: `http://tags.library.upenn.edu/`

However, the authors find that for many documents the proportion of each tag becomes more or less fixed as time passes, usually after about 100 users have tagged the document. This implies that after some time a consensus among the users on how the document should be categorised can usually be observed. While other users can still assign more tags to the document, the consensus is not likely to be changed.

A similar observation is made by Halpin et al. (2007). The authors propose a generative model of collaborative tagging and study how a stable distribution of tags is reached. By studying 500 URLs which have been tagged by more than 2,000 users, they confirm that the tag distributions always converge to a power law distribution, meaning that there are always a few popular tags which are used by most of the users and a 'long tail' of tags which are used by very few users. The authors also employ Kullback-Leibler Divergence, a measure of the distance between two probability distribution, to investigate the trend of tag distribution convergence. It is found that tag distribution usually converge to a power law distribution within several months for most of the URLs investigated.

Golder and Huberman (2006) suggest that there are two major reasons for this phenomenon, namely imitation and shared knowledge. While users are free to use any tags to describe a document, they are usually affected by the tags assigned by previous users. This is especially true in Delicious where common tags assigned to the document by other users are presented to a user when he/she first assigns tags to the document. In addition, users of Delicious probably share some cultural or background knowledge which makes them view the documents from a common perspective, and subsequently categorise the documents in a similar fashion.

This phenomenon of converging tag distribution is actually related to what is called *indexing consistency* in traditional subject indexing. Indexing consistency refers to the extent to which agreement exists on the terms to be used to index some documents among different indexers (Lancaster, 2003). This is commonly measured by calculating the overlap between the sets of indexing terms used by different indexers. According to Hooper (1965), a high level of consistency is very difficult to achieve. His summarisation of fourteen previous studies reveals that indexing consistency can range from 10% to 80%. This is also related to the vocabulary problem described by Furnas et al. (1987), who reveal that it is very common for users to use different words to refer to the same thing. It is suggested (Lancaster, 2003) that indexing consistency is dependent on several factors as listed in Table 2.3.

1. Number of terms assigned
2. Controlled vocabulary versus free text indexing
3. Size and specificity of vocabulary
4. Characteristics of subject matter and its terminology
5. Indexer factors
6. Tools available to indexer
7. Length of item to be indexed

TABLE 2.3: Seven possible factors affecting consistency of indexing (Lancaster, 2003).

While it is out of the scope of this thesis to look into each of these factors in detail, it is worthwhile to look at these dimensions in the context of collaborative tagging. Interestingly, collaborative tagging systems tend to be systems in which consistency is least expected. Collaborative tagging does not involve a controlled vocabulary. Allowing users to create tags on-the-fly means that the vocabulary used in a folksonomy can be huge in size and vary greatly in terms of specificity. Items being tagged generally span a wide range of topics and thus involve terminologies from different domains. Indexers are general Web users in collaborative tagging and they possess different level of expertise. In addition, there are various types of items of different lengths. All these factors seem to suggest that indexing consistency should be low in collaborative tagging.

One important point to note is that convergence in tag distribution is related but not equivalent to indexing consistency. Given the observed power law characteristic of stabilised tag distributions, the indexing consistency between users must be very low. For example, this is shown by Noll and Meinel (2008a) who study consensus of users by using entropy as a measure of agreement between users. High consistency would actually result in a more or less even tag distribution. However, a power law distribution suggests that users in general only agree with each other on the first few tags. This is actually similar to the situation described by Lancaster (2003) in which indexers tend to agree with each other on the first few index terms if they are asked to give terms in the order of their perceived importance. Although collaborative tagging in general does not assume an order of precedence in the tags supplied by the users (users cannot associate a weight with a particular tag when tagging a particular item), they collectively arrive at a consensus on which are the most important terms for an item, as evidently shown by the power law distribution.

Bischoff et al. (2008)	Golder and Huberman (2006)	Xu et al. (2006)	Sen et al. (2006)
Topic	What or Who it is about	Content-based	Factual
Time	Category Refinement	Context-based	
Location			
Type	What it is	Attribute	
Author/Owner	Who owns it		
Opinions/Qualities	Qualities/Characteristics	Subject	Subjective
Usage Context	Task organisation	Organisational	Personal
Self Reference	Self Reference		

TABLE 2.4: Mapping between different proposal of categories of tags (Bischoff et al., 2008).

2.3.2.4 Types of Tags

The construction and usage of tags, unlike names of topics or categories in traditional methods of categorisation or classification, are not restricted. A tag can be a single word or a short phrase, a noun or an adjective. It can also be an objective classification, or a subjective description. This provides users with the freedom to use tags of various kinds to suit their own purposes. Golder and Huberman (2006) identify seven functions of tags in Delicious, such as for indicating what or who the document is about, for indicating the type of the document, and for organising one's own tasks. These can also be seen as seven kinds of tags that are commonly used by the users of Delicious.

Xu et al. (2006) come up with a similar categorisation of tags by studying the collaborative tagging system of Yahoo! My Web 2.0.[8] The five categories described by the authors include: (1) content-based tags; (2) context-based tags; (3) attribute tags; (4) subjective tags; and (5) organisational tags. These five categories actually cover a set of similar types of tags found in collaborative tagging systems and represent a categorisation more general than that of Golder and Huberman. For example, tags for self reference and for task organising can be grouped under the category of organisational tags. Bischoff et al. (2008) further refine the categories proposed by Golder and Huberman (2006) by introducing the classes *Time* and *Location*, which are mainly found in, for example, Flickr when images are tagged. Table 2.4 presents the mappings between the different categorisations of tags mentioned above as described by Bischoff et al. (2008).

In fact, the types of tags appearing in a collaborative tagging system very much depend on the resources being tagged in the system. While Golder and Huberman (2006) and Xu et al. (2006) study collaborative tagging systems which involve

[8]Yahoo! My Web 2.0: `http://myweb2.search.yahoo.com/`

bookmarking URLs, other types of tags can actually be found in collaborative tagging systems which involve other types of resources. For example, in Flickr tags such as `bright` and `red` are commonly found to be used to describe digital images uploaded to the system. This type of tags which describes the visual features of an item is thus specific to collaborative tagging systems of images. Likewise, Bischoff et al. (2008) report that tags which represent genres of music account for a very large proportion of tags used in Last.fm, in which users mainly use tags to describe songs and music records.

It can be observed that tags are clearly not limited to indicating the topic of a document as in traditional subject indexing or categorisation. Tags are also used to describe something that are related to the document. Tags can also be used in a very personal and idiosyncratic way to suit the users' own purpose, such as indicating that a document is interesting or worth reading at a later time. In fact, one of the most noteworthy characteristics of collaborative tagging is the abundance of organisational or personal tags. Beside the common organisational tags such as `howto` and `toread` in Delicious, we also see tags such as `living room shelf` and `bedroom` in LibraryThing. This suggests that users are treating collaborative tagging systems as much a personal tool for information organisation as a social tool for sharing their favourite resources.

2.3.2.5 Network Analysis

As we have discussed in Section 2.3.1, a folksonomy is basically a set of associations between three basic entities—users, tags and documents. Therefore a folksonomy can always be represented in the form of a complex network, with nodes representing the entities and edges representing their association. As folksonomies involves three different types of nodes, the underlying networks are usually tripartite hypergraphs (Mika, 2007; Catutto et al., 2007; Lambiotte and Ausloos, 2006; Niwa et al., 2007).

By representing folksonomies as complex networks, network analysis becomes another means of understanding the different characteristics of folksonomies. As one of the earliest studies which analyse the structure of complex networks of folksonomies, Shen and Wu (2005) study whether the networks display characteristics of small world (Watts, 1999) or scale-free networks (Barabasi and Albert, 1999). They analyse data obtained from Delicious and construct a network of tags, in which tags are connected by edges if they have been assigned to the same document by some users. The authors discover that the network of tags features much

higher clustering coefficient than a random network generated by specifying the same number of nodes and edges, while having similar value of characteristic path length. This shows that the network of tags can be considered as a small world network. It is further shown that the degree distribution of the network follows a power law distribution, suggesting that the network is also a scale-free network. Nevertheless, the authors emphasise that based on these results the network of tags is said to be small world and scale-free only in a local sense, as the property of the whole network may be different from that of the fragment in their study.

Rather than studying only the network of tags obtained from a folksonomy, Catutto et al. (2007) consider the much larger tripartite hypergraph of a folksonomy. In particular, the small world characteristics of the folksonomies in Delicious and BibSonomy are examined. Since the folksonomy networks are tripartite hypergraphs which involve three disjoint sets of nodes instead of a single set of nodes as in the original studies of small world networks, it requires a different method for measuring its clustering coefficient. Based on the notion of clustering coefficient, the authors define two different but related coefficients to measure the degree of clustering in the tripartite networks. The first one is called cliquishness, which indicates the proportion of possible edges that actually exists around a certain node. The second one is named connectedness, which measures the extent to which the neighbours of a certain node are connected to each other. The experiments compare the folksonomy networks with random networks. The results show that the clustering coefficients of the folksonomy networks are very high and their characteristic path lengths are comparable or even smaller than those of the corresponding random graphs, indicating that the networks are small world networks.

These studies provide us with a better understanding of the structure of folksonomies. In particular, we understand that the underlying structure of folksonomies can be described by the small world or the scale-free models. One of the benefits of having this knowledge is that the future growth of a folksonomy can be predicted to a certain extent so as to guide the development of more efficient collaborative tagging systems. In addition, it also helps explain some of the phenomena of folksonomies. For example, as suggested by Catutto et al. (2007), the small world structure is probably why serendipitous browsing of Web resources is so popular among users of collaborative tagging systems. This is because even though the folksonomies growth in a rapid rate, the nodes are still reachable within a few clicks.

2.3.2.6 Efficiency

The simplicity of folksonomies encourages users to organise their favourite resources on the Web with tags. However, as the number of users, tags and resources in a collaborative tagging system continue to increase over time, the efficiency and scalability of the underlying folksonomy come into question. Instead of looking into aspects such as system architecture and design of a collaborative tagging system, it is more interesting to study whether a folksonomy will remain an efficient and useful tool as its size continues to grow. For example, when more and more documents are assigned the same tag, is the tag still useful for the purpose of organising, managing and sharing documents? Is it still easy for a user to retrieval documents of a particular topic? In view of these questions, Chi and Mytkowicz (2008) examine and analyse the usefulness and navigability of tags in collaborative tagging systems by employing information theory. By calculating the entropy and conditional entropy of tags they demonstrate a means of evaluating the effectiveness of tags.

As a core concept in information theory (Shannon, 1949), ***entropy*** measures the average amount of information associated with a certain random variable. In other words, it gives an idea of how predictable the outcome of the random variable is. A higher entropy means the outcome is less predictable. The entropy of a random variable basically depends on two aspects. Firstly, the entropy will be higher if the number of possible events is larger. Secondly, the entropy will also be higher if the probability distribution of the random variable is more uniform.

In the context of collaborative tagging, entropy can be used as a measure of the effectiveness of tags in a folksonomy in encoding documents. For example, if we consider T as a random variable with all the tags in the system as its possible events, the entropy of T, denoted $H(T)$, gives us an idea of the diversity of the tags used in the folksonomy. Based on a dataset obtained from Delicious, Chi and Mytkowicz (2008) show that $H(T)$ first increases and starts to plateau after some time. Given that the number of bookmarks keeps increasing over that period of time, the authors suggest that the plateau of $H(T)$ means that existing and popular tags in the folksonomy are used again and again, resulting in a much less uniform distribution of tags, which compensates for the effect of increasing number of tags. This in fact agrees with the observation that tag distribution usually settles to a power law distribution, as we have discussed in Section 2.3.2.3 (power law distributions have very low entropy values due to the high probability of a few possible outcomes). In addition, the authors also find out that the entropy

measure $H(D|T)$ increases over time. This suggests that the number of documents assigned a particular tag keeps increasing, meaning that retrieval using tags will return more and more documents and that the users will need to spend more effort in browsing through the list for documents which satisfy their needs.

The above study provides only a partial view of the problem as it considers only the efficiency of a single tag in annotating a single document. However, it does provide insight into a major issue of folksonomies: how can we achieve a balance between efficient use of tags (reducing $H(T)$) and efficient annotation of documents (reducing $H(D|T)$). This is relevant to both the whole tagging community as well as to individual users. A possible solution to this problem of reduced efficiency is to employ ranking algorithms when retrieving documents from a folksonomy, which present a list of documents ranked according to their relevance to the specified tags. We will discuss more about information retrieval and ranking in the context of collaborative tagging in Section 2.3.3.

2.3.2.7 Spamming and Folksonomies

While general Web users find collaborative tagging a good method for organising their resources and sharing them with their friends, malicious users have also been increasingly attracted to the popularity of collaborative tagging systems. In particular, as more and more users use collaborative tagging systems to search for new resources on the Web, these systems become new arenas for spammers to bring their content to the attention of users. Moreover, as most collaborative tagging systems provide convenient interfaces for tagging, spammers can easily automate their tasks. Although collaborative tagging is still a relatively new phenomenon, spamming activities can already be observed in popular systems. For example, Wetzker et al. (2008) find out that 19 out of the 20 most active users (who have the largest number of bookmarks) in Delicious are actually spammers.

Heymann et al. (2007) discuss spamming in social Web sites in general with emphasis on collaborative tagging systems. The authors describe three different types of anti-spam strategies for protecting a collaborative tagging system from spamming activities, namely detection, demotion and prevention. Detection measures aim at identifying spammers and spams (unwanted resources) by studying the tagging history within a system to identify abnormal behaviour. Regarding this type of measures, a number of papers from the Discovery Challenge in the ECML/PKDD 2008 Conference provide several solutions based on machine learning techniques.

For example Kim and Hwang (2008) propose the use of a naive Bayes classifier for learning the behaviour pattern of spammers for spam detection, while Gkanogiannis and Kalamboukis (2008) adapt a text classification algorithm for the purpose. Neubauer and Obermayer (2009) approach the problem of spam detection by constructing hyperincident networks, in which vertices represent hyperedges in the folksonomy graph. The authors discover that hyperincident networks that are polluted by spams feature connected components that exhibit characteristics very different from those that are not affected by spam.

Demotion measures, on the other hand, try to reduce the prominence of spams in a collaborative tagging system by giving them lower ranks compared to other legitimate content. This type of measures is particularly useful when users are presented lists of resources in a tagging system. Instead of identifying the spammers, demotion measures try to hide the spammers from the reach of users as much as possible, thus reducing their negative effect on the system. Koutrika et al. (2008) suggests that frequency-based ranking methods are vulnerable to spams. The authors propose the 'coincidence factor', which measures the 'reliability' of users who have assigned tags to resources, as a ranking method to demote spammers who deliberately assign wrong tags to resources for different malicious purposes.

Finally, prevention-based anti-spam measures try to avoid malicious users to use the systems in the first place. A tagging system can challenge the user with some problems which can easily be solved by a human user but appear more difficult to computer programs. In this way spammers will find it more difficult to deploy automated bots to deliver spams to the system. For example, CAPTCHA (Completely Automated Public Turing Test for Telling Computers and Human Apart) (Ahn et al., 2003), which challenges users using hard AI problems, is a possible way of implementing prevention-based measures.

Anti-spam is an important issue of collaborative tagging and has significant implications. If a collaborative tagging system is heavily spammed by unwanted materials, its value in helping users to organise and share online resources is significantly decreased. In addition, research on the derived applications of folksonomies must also take spamming into account, as spamming activities can be so common that there can be a lot of noise in the data. For example, the outcomes of using a folksonomy to bootstrap semantic applications or to construct recommendation systems may be affected by the presence of spams in the folksonomy.

Although each type of the above measures provides some solutions to the problem, they all have their limitations. Detection measures very often require a large set of

data for learning the behavioural pattern of spammers. Demotion-based ranking algorithms are susceptible to gaming by some focused spammers. Prevention-based measures may reduce the incentive of ordinary users in using the systems. Hence, it is likely that a well-balanced combination of these three kinds of measures is needed. A more thorough understanding of the spamming activities in collaborative tagging system is also required.

2.3.3 Folksonomies and Information Retrieval

The popularity of collaborative tagging leads to a large number of user-contributed annotations of documents available on the Web. It is therefore natural to ask whether this kind of new information—keywords that are assigned by the readers of the documents and were not available to Web search engines before—is able to contribute to improving search and retrieval of documents on the Web. Indeed, we see quite a lot of studies in the literature investigating different issues in this area. In this section we give a summary of these studies and discuss their findings.

2.3.3.1 Metadata and Web Search

Metadata of documents refer to data that describe something about the documents, such as what the document are about, who the creators of the documents are and when they are created or modified. In the context of the Web, the most common method for the author of a Web document to specify metadata is by using the `META` element defined in the HTML standard.[9]. The `META` element allows the author of a document to specify a set of keywords and even a detailed description of the document's content. These metadata are intended to facilitate retrieval of the documents.

Generally speaking, there are three types of metadata (NISO, 2004). Firstly, *structural metadata* represent information used to describe the nature of a document. For example, this can be the file type of the document, or the size of the document. Secondly, *administrative metadata* represent information describing the handling of the document. These include for example the publication date, the date of last changes made, the current status of the document, etc. Finally, there are *descriptive metadata* which tell what the document is about. This type of metadata is actually the type of metadata that for most people are the natural entry

[9]See `http://www.w3.org/MarkUp/`

point to large volumes of information. For example, Web users usually search for documents that address certain topics instead of for documents of certain types.

As we have discussed in Section 2.3.2.4, tags are used to describe documents in many different ways. Therefore tags can be treated as a type of metadata of the documents to which the tags are assigned. Tags actually cover all the three different types of metadata that we have just described. Unlike metadata specified by the author of a Web document in the META element, tags are assigned by users and are therefore metadata that describe the document from the users' perspective. It is likely that there are differences between the two. It is therefore interesting to investigate whether tags can be considered as a good form of metadata and whether they can be harnessed to enhance Web search.

One focus of research works in this direction is whether tags provide new information about the documents to which they are assigned. Noll and Meinel (2007b) reveal that tags assigned by users can usually be found in the HTML metadata or the body of the document (58.4% of the tags can be found in either the HTML metadata or the body text of the documents). In a related study (Noll and Meinel, 2008b), the authors measure the amount of new information provided by several different types of metadata, namely anchor text, search queries and tags, by using a measure called *novelty*, which refers to the percentage of unique terms that are not present in the document. They reveal that in general tags are likely to be terms which can be found within the documents. Heymann et al. (2008a) also report similar findings on their datasets, revealing that tags are present in the body text of 50% of the documents they are assigned to). While this suggests that there is a significant overlap between the information provided by tags and that contained in the documents, a simple conclusion as to whether tags can be used to enhance Web search can not be made based on these findings. What kind of new information do the tags provide? How is this information related to the queries often used by users in search engines? Answers to these questions are yet to be found out.

Other aspects in this research direction include the quality and quantity of tags. While Delicious is very popular among Web users nowadays, it is estimated that the number of URLs covered by Delicious is still very small compared to the number of documents available on the Web (Heymann et al., 2008a). Bao et al. (2007) also mention that using tags to enhance Web search is restricted by the limited number of tagged documents. However, Noll and Meinel (2007b) and Heymann et al. (2008a) also report that documents which are assigned tags in

Delicious tend to be popular documents (as indicated by their PageRank value and their appearance in Web search results). While it is not clear whether users tend to tag popular documents or tagged documents tend to become popular, this suggests that using tags in Web search may help to locate documents of higher quality. In addition, several studies (Brooks and Montanez, 2006; Noll and Meinel, 2007b, 2008a) reveal that tags usually represent broad categories instead of specific ones, suggesting that tags may be more suitable for grouping documents into general categories than helping to locate documents of a specific topic.

Some studies suggest that tags are limited in their usefulness in enhancing Web search in one way or another. However, there are also studies that successfully utilise tags to providing better search results (Bao et al., 2007; Noll and Meinel, 2007a; Yanbe et al., 2007). We believe that the main point of this discussion does not lie in the question of whether searching by using tags alone would produce high quality results. Instead, the challenge is how tags can be utilised to improve Web search results whenever they are available. For one thing, that tags do not contribute new information to the metadata of a document does not mean that they are less useful. It must be noted that that a keyword is present in the body text of a document does not necessarily mean that it would be selected to be included in the metadata. However, tags are keywords that are already singled out by the users to describe the document. It is expected that when more about the interactions between tags, queries and other types of metadata is understood, tags can be harnessed to improve Web search in different aspects.

2.3.3.2 Ranking in Folksonomies

Information retrieval and Web search do not only focus on whether relevant and high quality documents can be retrieved, but also focus on whether the most relevant documents can be presented to the users first (Manning et al., 2008). This is also true in collaborative tagging as the number of documents being tagged keeps increasingly rapidly in many systems. Folksonomies pose a more complicated data structure for the task of ranking. In traditional information retrieval, it is always the documents which are being ranked. However, in the tripartite structure of folksonomies, it is equally reasonable and desirable to ask for a ranking of users, tags or documents.

Hotho et al. (2006a) are among the first to study ranking in folksonomies. The authors propose the FolkRank algorithm, which is an adaptation of the PageRank algorithm (Brin and Page, 1998), for ranking entities in a folksonomy. The

algorithm is a topic-specific and personalised ranking method that makes use of a preference vector. John and Seligmann (2006) propose ExpertRank for ranking users of their expertise in a particular topic represented by a tag. The algorithm ranks users according not only to how many times they have used a particular tag but also gives higher scores to users if they have also used related tags determined by co-occurrence analysis. Similarly, Bao et al. (2007) propose the SocialPageRank algorithm based on the mutual reinforcement of the levels of popularity between the three entities in a folksonomy. This can be considered as an adaptation of the HITS algorithm (Kleinberg, 1999) to the tripartite structure of a folksonomy.

All these algorithms are based on the assumption that important/popular resources are likely to be assigned important/popular tags by important/active users. However, this also means that rankings produced by these algorithms tend to put too much emphasis on the popularity of a tag or a document, or the activeness of a user. As we have discussed in Section 2.3.2.7, popularity and activeness are not as reliable as they seem due to the presence of spamming activities. We will discuss more about the relations between ranking and spamming activities in the context of collaborative tagging in Chapter 5.

2.3.4 Folksonomies and Recommendation Systems

Given their collaborative nature, tagging systems are usually compared with, or are considered as a new form of collaborative filtering (Marlow et al., 2006). Collaborative filtering (Herlocker et al., 2004) is a method of making predictions about the interests of users by identifying users having similar histories on certain activities. The information is then used to recommend new items to users. Collaborative tagging systems, in which users express their own opinions on the classification of Web resources, can also be considered as a kind of collaborative filtering in which tags act as votes. Users who use similar tags or are interested in similar resources are likely to share same interests. How collaborative tagging can be exploited for recommendation systems (Montaner et al., 2003) is thus one of the current research directions.

A prerequisite for recommendation systems to work effectively is that they must have good understanding of the interests of the users. The task of modelling user interests is commonly known as user profiling (Godoy and Amandi, 2005). A major issue in user profiling is how accurate information about user interests can be obtained. While collaborative tagging systems may not have any thing to do with

user profiling at first glance, the tags and documents posted to the systems by the users actually provide a lot of information for understanding their interests. In particular, Li et al. (2008) find out that tags are in general more appropriate than term weighting schemes such as TF-TDF in representing human being's judgements about Web content, and are thus good candidates for representing user interests.

In the simplest form, a user profile constructed from a folksonomy takes the form of a tag vector, in which each element reflects how many times the corresponding tag has been used by the user. For example, Diederich and Iofciu (2006) propose TBProfile, which is a user profile generator based on the annotations assigned to the documents interested by a user. Similarly, Noll and Meinel (2007a) use a tag vector as a user profile to provide personalised Web search based on data in a folksonomy. Michlmayr and Cayzer (2007) present some slightly more sophisticated methods for constructing user profiles. These include selecting the top k pairs of tags which are most frequently used together, or using a time decay function to modify the weights of the tags in a user profile to reflect the interests of the user over time.

A single tag vector is however not sufficient to model user interests accurately in many cases. In particular, users are usually found to be interested in diverse topics (Kook, 2005). One limitation in using single tag vector is that, similarity between users will tend to be low when users have multiple interests because a single tag vector collapse the differences between these interests when performing similarity calculation. In traditional user profiling studies, quite a number of authors have proposed the use of multiple term vectors to accommodate the multiple interests of users (Segal and Kephart, 2000; Chen and Sycara, 1998; Pon et al., 2007). Our study (Au Yeung et al., 2009b) on the usage of tags of individual users in Delicious confirms that they do have diverse interests. We also propose an algorithm for discovering and modelling multiple interests of a user based on cluster analysis of his/her personomy.[10]

In addition to user profiling, there are quite a number of studies in the literature which focus on exploiting folksonomies for resource recommendation. Niwa et al. (2006) propose a Web page recommendation system based on data obtained from Delicious. The system abstracts each user's interests by associating each user with a tag cluster, and thus pages that are more related to the tag clusters can be recommended to the users. Shiratsuchi et al. (2006) also propose an information

[10] A personomy refers to the whole collection of tags, resources and tag assignments of a single user (Hotho et al., 2006a).

recommendation system based on folksonomies. The authors use a clustering algorithm to extract communities of users from a network of users constructed from a folksonomy. Resources of a user can be recommended to other users within the same community. Shepitsen et al. (2008) also describe a similar system for personalised recommendation by using agglomerative hierarchical clustering to discover sets of tags of different topics.

All the above studies and proposals suggest that folksonomies contain much information about user interests, and they can be exploited for the purpose of recommending potentially interesting resources to users. In fact, one of the advantages of using folksonomies for this purpose is that both the interests of the users and the topics of the resources have the same form of representation: they are all characterised by tags created by users. However, performance of recommendation systems built on top of collaborative tagging can be limited by problems of folksonomies such as the existence of synonymous and ambiguous tags. In addition, the density of the underlying folksonomies may affect the accuracy of recommendation (Shepitsen et al., 2008). In this respect, Szomszor et al. (2008a) propose methods to construct more comprehensive user profiles by combining the tags used by same users in different folksonomies (e.g. Delicious and Flickr).

2.3.5 Tag Clustering and Co-occurrence Analysis

As one would notice from the above discussions, tag clustering—grouping tags that are highly associated with each other—is a major process of many analyses and derived applications of folksonomies. The process involves finding semantically related tags based on some similarity measures that exploit the associations between different entities in a folksonomy.

Tag clustering is not a trivial process because there are many different ways to model a folksonomy (or a subset of a folksonomy) and to measure similarity between tags. The simplest form of tag clustering is by constructing a tag co-occurrence network. Such network is constructed by treating tags as nodes and using edges to connect two nodes if the tags have been assigned to, or co-occur in, the same document. As pointed out by Begelman et al. (2006), the use of absolute numbers to represent the strength of tag relations will make the cluster analysis heavily bias to popular tags. Therefore they propose the use of some similarity measures such as the Dice coefficient or the Jaccard coefficient to calculate the relative strength of tag relations before clustering algorithms are applied. The au-

thors perform clustering of such a network of tags constructed with data obtained from Delicious by using spectral clustering algorithms. They demonstrate that tag clustering can be used to group semantically related tags together.

Brooks and Montanez (2006) perform analysis on data collected from Technorati, a search engine of blogs in which users can use tags to categorise blogs.[11] They tackle the problem of tag clustering in an indirect way by first performing clustering on documents being tagged. The authors collect a set of popular tags from the Web site, and for each of the tags a set of most recent documents are collected. Agglomerative clustering techniques are then applied to construct a hierarchical structure of the clusters of documents, with a similarity measure based on the vector space model and the TF-IDF term weighting scheme. It is shown that tags with are semantically related to each other are clustered together.

Mika (2007) investigates two different ways of constructing tag networks for the purpose of lightweight ontology generation. On the one hand, a *community-based tag network* is constructed by connecting tags that have been used together by the same user. On the other hand, an *item-based tag network* is constructed by connecting tags that have been used together on the same document. The author presents two case studies, one of the tags obtained from Delicious, and another of the terms extracted from Web pages. The community-based tag network is found to provide a clearer and more precise picture of the relations between the tags: in the Delicious case, the clustered tags reflect the core interests of the users, while in the latter case feedbacks from members of the community indicate that a majority of the members consider the community-based network as more accurate in reflecting relations between the terms extracted. This work, which first studies tag clustering in the social context, suggests that semantic relations between tags are inseparable from the context of the community in which they are created or used.

Cattuto et al. (2008b) present an in-depth study of relatedness of tags in Delicious. The authors compare several different similarity measures of tags, including simple co-occurrence, tag/user/resource context similarity, and similarity based on the FolkRank (Hotho et al., 2006a) algorithm. It is found that while each similarity measure returns some meaningful results, the types of similarity they account for can be quite different. In particular, it is found that tag context similarity and resource context similarity both tend to return pairs of tags that are synonymous to each other (semantically similar), while measures such

[11] Technorati: `http://www.technorati.com/`

as co-occurrence and FolkRank return pairs of related tags. For example, for the tag `opensource`, the most similar tags according to tag context similarity are `open_source`, `open-source` and `open.source`, whereas those according to co-occurrence are `software`, `linux` and `programming`. The author therefore suggests that different similarity measures should be chosen in different application scenarios.

The associations in a folksonomy can also be exploited by using some traditional data mining techniques. For example, Schmitz et al. (2006) propose finding implicit dependencies between entities in a folksonomy by using association rule mining (Agrawal et al., 1993), which is a popular data mining technique for discovering potentially useful relations between items in a large database. The authors present methods for projecting the tripartite structure of a folksonomy onto some two dimensional space for mining rules for a particular entity. Heymann et al. (2008b) also the technique to mine association rules of tags (e.g. users assigning tag T_1 to some resources are likely to assign T_2 as well) to predict which tags are likely to be assigned to a particular resource. In a broader sense, association rule mining in folksonomies is related to the construction of subsumption hierarchies of tags and generation of lightweight ontologies. This will be discussed in more details in Section 2.3.7.

2.3.6 Synonymy and Ambiguity

Synonymous and ambiguous tags are inevitably common in folksonomies due to their unrestricted nature. As mentioned in Section 2.3.2.2, the existence of these tags is believed to be limiting the effectiveness of collaborative tagging in organising and retrieving resources. Hence, effective methods for identifying synonymous tags for a particular tag and for identifying the different meanings of an ambiguous tag are desirable.

While in general synonymous tags can be found in the same cluster in tag clustering, it is not a trivial task to distinguish between synonymous tags and related tags. Niwa et al. (2007) are the first to present a targeted solution to the problem of identifying synonyms. The authors propose a method for estimating the relationships between tags by calculating both document-based and user-based tag co-occurrence. They introduce a heuristics to discover synonymous tags as follows: tags that are synonymous to each other are usually assigned to the *same* document by *different* users. For example, a user who has assigned `semweb` to a

document is less likely to assign `semanticweb` to the document at the same time, but it is likely that other users have assigned `semanticweb` but not `semweb`. They implement this heuristics by calculating both user-based mutual information and document-based mutual information. It is reported that this method achieves a 92% accuracy on data obtained from Delicious. Clements et al. (2008) present a similar approach to identifying synonyms in LibraryThing, implementing the above heuristics using the Pearson correlation measure.

It should be noted that the usefulness of such heuristics is also dependent on the design of a collaborative tagging system. For example, Delicious offers tag suggestions when a user adds a bookmark to the system. A user may therefore be encouraged to add synonymous tags suggested by Delicious to help retrieval of the resource being tagged if he/she aims at sharing it with other users. If more users are adding synonyms in this way, the assumption of the above heuristics will be weakened. In this respect, the tag context similarity or resource context similarity studied by Cattuto et al. (2008b) may be useful for complementing this heuristics to identify synonyms.

In contrast to the problem of synonyms, few studies in the literature actually focus on the problem of discovering the different meanings of ambiguous tags, although quite a number of authors have acknowledged the existence and impact of these tags. In general, these tags are tags that possess different meanings or represent different concepts when used in different contexts. Here we mention some related works in this area. We will discuss the nature of these tags in detail in the next chapter.

Wu et al. (2006) explore the possibility of deriving emergent semantics (Aberer et al., 2004) from social annotations by mapping tags onto a high dimension vector space. Using a probabilistic model, the authors estimate the probability of a tag appearing in each of the dimensions which represent different categories of knowledge. The basic idea is that the number of dimensions in which a tag scores high correspond roughly to the number of meanings the tag has. Based on the probability distribution of a tag on the chosen dimensions, the ambiguity of the tag can be characterised by the entropy of the tag (a high entropy corresponds to high ambiguity). One drawback of this approach to identifying ambiguous tags is that one needs to choose the number of dimensions beforehand. While the log-likelihood as suggested by the authors can be used as a reference of how many dimensions should be chosen, the number also depends on the amount as well as the diversity of tags being processed. The number of dimensions may limit the

number of meanings which can be discovered for a tag.

Zhou et al. (2007b) propose using deterministic annealing together with a divisive hierarchical clustering algorithm to derive hierarchical structures of tags in Delicious and Flickr. The authors report that tags which are ambiguous can be found appearing in different branches in the resultant hierarchy, thus indirectly identifying the multiple meanings of these tags. However, in this work the identification ambiguous tags is more of a by-product of the hierarchical clustering process, and the method cannot be used directly to identify the multiple meanings of a particular tag.

2.3.7 Folksonomies and Ontologies

As we have discussed in Section 2.3.2, folksonomies are always compared with some more formal knowledge representation methods such as taxonomies and ontologies. Since construction of ontologies in general requires much effort and a consensus on the vocabulary is difficult to reach in many domains due to for example the large number of users or the large number of concepts and relations involved. Hence, there has been wide discussion on whether folksonomies can be used as a bottom-up approach to generate ontologies by harnessing the collective effort of a community of users.

A research problem commonly mentioned in this respect is one of generating a hierarchical structure of tags from a folksonomy. The hierarchical structure usually represents a set of subsumption relations between tags. As discussed in the previous sections, Zhou et al. (2007b) use a divisive hierarchical clustering algorithm to generate hierarchies of tags in Delicious and Flickr. Schmitz (2006) makes use of a probabilistic model to identify subsumption relations between tags in Flickr. Association rule mining has also been used to identify rules which can be considered as subsumption relations between tags (Schmitz et al., 2006). These techniques, however, are usually limited to finding subsumption relations in the broad sense: that if tag A is used then it is likely that tag B will also be used. The semantics of the relations between tags is therefore less likely to be reflected in the hierarchical structures generated by these techniques.

Van Damme et al. (2007) describe a more comprehensive set of methods for deriving ontologies from folksonomies by combining different techniques of co-occurrence analysis and a wide range of resources available on the Web. For example, the authors suggest mapping tags in folksonomies to entities in Wikipedia,

using WordNet (Miller, 1995) to identify synonyms and homonyms and using the Google search engine to check spellings of tags. Specia and Motta (2007), along the same research direction, propose a system for adding explicit semantics to tags by combining statistical analysis of folksonomies and knowledge provided by existing ontologies available on the Web. While these kinds of techniques generate results that provide more semantic information about the tags in a folksonomy, there are difficulties in applying them widely to a large number of tags because tags may address a wide range of topics that existing lexical resources or ontologies.

Rather than trying to generate taxonomies or ontologies from folksonomies, Gruber (2007) suggests that ontologies and in general Semantic Web technologies can be used to enhance the value of folksonomies. He suggests using ontologies to provide a formal conceptualisation of the activity of tagging, thus allowing user-contributed metadata in different folksonomies to be combined together for better retrieval of resources across folksonomies. He proposes formalising the activity of tagging as a five-place relation involving an object, a tag, a tagger, a source and a binary vote ([+/−]). The object is the Web resources being tagged, the tagger is the user who assigns tags, the source refers to the system from which this annotation originates (for example this can be Delicious or Flickr), and [+/−] represents either a positive or negative vote placed on this annotation by the tagger (thus allowing a user to say that a particular tag is not applicable to the object). Newman (2004c) also proposes a similar ontology for collaborative tagging using RDF (the Resource Description Framework).[12] In the ontology a class called 'Tagging' is defined to bind a user, a set of tags and an object together.

Other similar examples of ontologies of collaborative tagging include the SCOT (Social Semantic Cloud of Tags) project (Kim et al., 2007) and the MOAT (Meaning Of A Tag) project (Passant and Laublet, 2008). SCOT defines classes and relations for the representation of the tag cloud of a particular user, i.e. all tags used by the user as well as their frequencies. On the other hand, MOAT provides properties for specifying the meaning of a tag by using URIs of Semantic Web resources. Both ontologies are specified using RDF and OWL. A thorough review of these ontologies and related works can be found in the survey paper by Kim et al. (2008).

[12]Resource Description Framework (RDF): http://www.w3.org/RDF/

2.4 Chapter Summary

Collaborative tagging and folksonomies open up a lot of research opportunities in a wide range of areas. We have seen a lot of research works focusing on exploiting folksonomies as some user-generated classification schemes to enhance search and retrieval of resources on the Web, we have also seen that folksonomies offer a lot of data for studying the user interactions on the Web, as well as the use of vocabulary within a community of users.

Given the background of research we have discussed in this chapter, we see that the research challenge lies in how we can make sense of the user-generated data in collaborative tagging systems so that we can make use of the knowledge to facilitate organisation and retrieval of information. We believe that the collective user behaviours and the implicit interactions between users must be analysed before such knowledge can be acquired, because data mining and analysis of the user-generated data must rely on the implicit relations between the entities in a system, and these relations are generated by the users themselves. However, with the exception of a few studies (e.g. the work by Mika (2007)), we can see that this kind of study is still rare in the literature, thus providing the motivation for the research work described in this thesis.

In the next chapter we will describe the primary subject of study in this thesis, Delicious, which is one of the most popular and earliest collaborative tagging systems on the Web. We will also describe our process of collecting research data from Delicious as well as the datasets collected.

Chapter 3

Delicious: A Collaborative Tagging System

As we have mentioned in the previous chapter, collaborative tagging has become very popular in recent years. Not only has there been a large number of collaborative tagging systems on the Web, many existing Web sites that involve organising and/or sharing resources on the Web have also started to provide similar functionalities. In carrying out the research work described in this thesis, we need a large amount of real world data that can be subjected to different analyses. It is most desirable that the collected data would be general enough such that conclusions can be generalised to other scenarios. This means that we would like to focus on a collaborative tagging system that covers a range of domains as broadly as possible. In this thesis, we choose to focus on the probably most popular and one of the earliest examples of collaborative tagging system, Delicious. In this chapter we give an introduction of Delicious, discuss why it is a suitable target of study, detail the process by which we collect data from Delicious, and give an overall description of the data sets we have collected and will be used in studies presented in the later chapters.

3.1 Introduction to Delicious

Delicious was launched by its founder Joshua Schachter in November 2005, when it first started to provide online social bookmarking services to users free of charge. While the content contributed by the users can be browsed by any Web users by accessing the Web site of Delicious, a user must register for a free account before

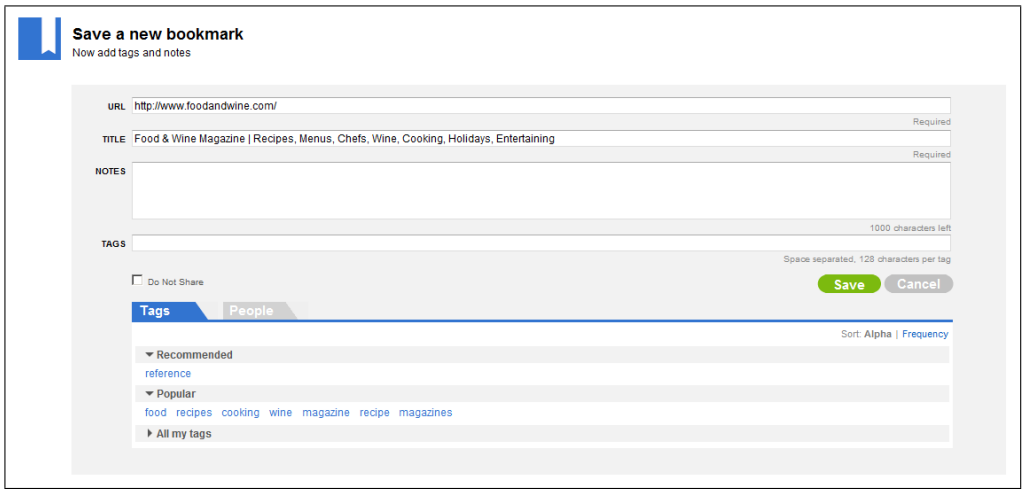


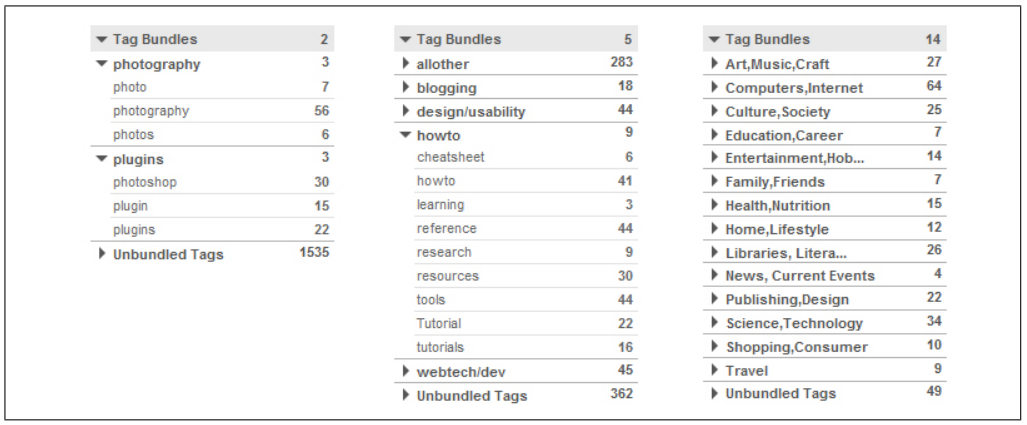
FIGURE 3.1: The interface for saving a bookmark to Delicious.

he/she can post bookmarks to the system. After registering with Delicious, a user can save a bookmark to Delicious by submitting a document (identified by a URL) to the system. At the same time, the user can specify a set of tags and also type in a comment to further describe the document. While the title of the document is automatically retrieved, the user can still modify the title if he/she wishes to do so. The pieces of information, including the URL, the title, any comments, and a set of tags, submitted by a user constitute a ***bookmark***. Furthermore, while users can supply any tags they like, tags which are commonly used to describe the document by other users in Delicious are suggested to the user in the bookmarking process. Figure 3.1 shows the interface for saving a bookmark to Delicious.

3.1.1 Organising and Sharing

Beside the basic functionality of saving bookmarks and assigning tags, Delicious has develop over the course of time quite a number of features to facilitate users to organise their bookmarks as well as their tags. The function of 'tag bundles' allows users to group their otherwise independent tags together under a common theme (see Figure 3.2).[1] For example, a user can create a tag bundle named `pets`, and then associate the tags `dogs`, `cats` and `fish` with this tag bundle. Retrieving documents in Delicious using this tag bundle will then resulted in a list of documents that have been assigned any of the three tags. In addition, users can also add a more detailed description to the tags they have used. These descriptions give better ideas on what the tags actually mean to the users and also remind the users of how their tags are used.

[1]Delicious Tag Bundles: `http://delicious.com/help/faq#tags`



Tag Bundles	2
photography	3
photo	7
photography	56
photos	6
plugins	3
photoshop	30
plugin	15
plugins	22
Unbundled Tags	1535

Tag Bundles	5
allother	283
blogging	18
design/usability	44
howto	9
cheatsheet	6
howto	41
learning	3
reference	44
research	9
resources	30
tools	44
Tutorial	22
tutorials	16
webtech/dev	45
Unbundled Tags	362

Tag Bundles	14
Art,Music,Craft	27
Computers,Internet	64
Culture,Society	25
Education,Career	7
Entertainment,Hob...	14
Family,Friends	7
Health,Nutrition	15
Home,Lifestyle	12
Libraries, Litera...	26
News, Current Events	4
Publishing,Design	22
Science,Technology	34
Shopping,Consumer	10
Travel	9
Unbundled Tags	49

FIGURE 3.2: Tag bundles of three different users in Delicious.

Delicious also provides several features for users to interact with each other and to allow them to share their bookmarks. At the most simplest form, a user can share a bookmark with another user by using tags of a particular format. Suppose user A would like to share a document with user B (with the user name 'userB'), user A can assign the tag `for:userB` to the document, and user B will be notified of this recommendation. In addition, if a user finds the collection of bookmarks of another user interesting and useful, he/she can add that user into his/her own network. Consequently every time that user has posted some new bookmarks to Delicious, the first user will be notified of the update. The first user then becomes a fan of the second user. Users can further group other users in their network into something called network bundles, such that users can be distinguished from each other according to their roles or their relationship with their fans.

3.1.2 Retrieving Tagged Resources

Retrieving documents in Delicious, obviously, relies primarily on tags. A list of documents that have been assigned, say, the tag `wine`, can easily be retrieved by accessing an URL of the form `http://delicious.com/tag/wine`. The list is ordered in reversed chronological order of the time at which the documents were last bookmarked by some users. Figure 3.3 shows an example of the interface. This method can be generalised to retrieving documents that have been assigned multiple tags, such as `wine`, `alcohol` and `shopping`, by using the plus sign to indicate conjunction: `http://delicious.com/tag/wine+alcohol+shopping`.

Beside the above method, Delicious also provides other different interfaces for users to browse data in the system from different perspectives. It is possible to browse a list of documents posted to the system by a particular user by accessing an URL

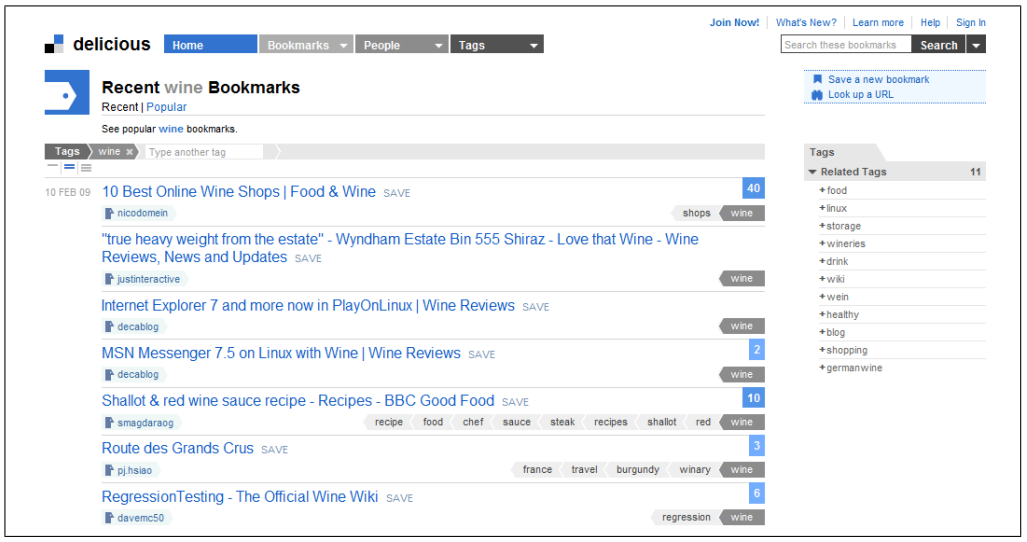


FIGURE 3.3: A list of documents assigned the tag `wine` on Delicious.

of the form `http://delicious.com/userA`. It is also possible to browse the whole bookmarking history of a particular document, i.e. a complete list of which user assigned which tag at which time (see Figure 3.4). Moreover, one can obtain a list of documents that have been assigned a particular tag (or a particular set of tags) by a particular user, using URLs of the form `http://delicious.com/userA/tag1`. Finally, the function of searching using keywords instead of only tags is also available. While the exact mechanism of this search function is unknown to the public, it is speculated that documents will appear in the result if their titles or descriptions contain the keyword(s) in the query.

3.2 Delicious as a Data Source

While there are many collaborative tagging systems and systems that provide collaborative tagging functionalities on the Web nowadays (e.g. LibraryThing, Bibsonomy and Last.fm), our analysis and studies are mainly carried out on data sets collected from Delicious. We have several reasons of why Delicious is chosen for the work described in this thesis. We elaborate each of these points as follows.

- **Popularity** Delicious is by far the most popular collaborative tagging system serving a broad folksonomy (see the next point) on the Web. As in July 2008, it was reported that Delicious was serving over 5 million users.[2] Having a large number of users means that data collected from the system

[2]See `http://blog.delicious.com/blog/2008/07/oh-happy-day.html`.

would cover a wider range in different dimensions. It also means that the chance of having an overlap between the different groups of users tagging a selected set of documents would be higher, thus providing us with a better basis of user behaviour analysis.

- **Data Model** As we have mentioned in Section 2.2.1, there are mainly two types of folksonomies, namely broad folksonomies and narrow folksonomies. The data model of Delicious is a prominent example of broad folksonomies, meaning that every user on Delicious is allowed to maintain their own sets of tags for the documents posted to the system. This allows us to look into the tags and documents with respect to the users at a more fine-grained level. This also means that the results of our analysis will be general enough to be applied to the tripartite structures of folksonomies in other collaborative tagging systems.

- **Interface** Delicious provides a very efficient user interface for users to browse data in the system from different perspectives. It is possible to browse a list of documents posted to the system by specifying a particular tag (see Figure 3.3) or a particular user name. In addition, it is also possible to browse the whole bookmarking history of a particular document, i.e. a complete list of which user assigned which tag at which time (see Figure 3.4). Such user interface greatly facilitates data collection as we can target our collecting process by specifying a particular tag, user or document.

- **Topics** Delicious is a general social bookmarking system that allows users to bookmark any resources on the Web regardless of the types and topics of the resources. While many authors (Golder and Huberman, 2006; Mika, 2007) report that many tags in Delicious are about Web and computing technologies as shown evidently by the list of popular tags, Delicious still represents a much more diverse platform when compared with other popular collaborative tagging systems, such as Last.fm and LibraryThing. Users in Delicious are also found to have a wider range of interests than users in other systems that focuses on a specific resource type (Bischoff et al., 2008; Au Yeung et al., 2008b).

In fact, a majority of studies about collaborative tagging in the literature use data collected from Delicious, as we have discussed in Chapter 2. We believe that the wide range of topics interested by the users of Delicious as well as Delicious' popularity will allow us to obtain results that can be generalised to other collaborative tagging systems, and even other social interactions on the Web.

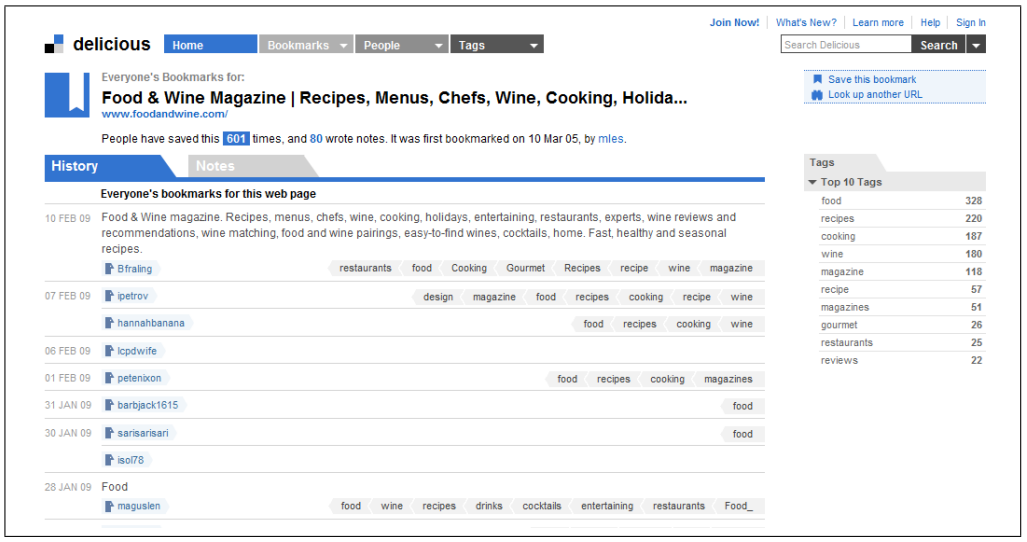


FIGURE 3.4: The bookmarking history of a particular document on Delicious.

3.3 Data Collection from Delicious

While we have mentioned that Delicious provides efficient interfaces to the data stored in the system, collecting data from Delicious is still not a trivial task. Beside the interfaces we have described in the previous section, Delicious also provides several different methods for accessing the collaborative tagging data. Firstly, users can subscribe to the RSS feed of a user, a tag or a document in Delicious in order to monitor updates of the corresponding entity. For example, the RSS feed of documents that are assigned the tag `wine` can be accessed at `http://feeds.delicious.com/v2/rss/tag/wine`. These feeds are presented in JSON (JavaScript Object Notation) or XML (eXtensible Markup Language) formats and can be easily parsed to obtain the data required for analysis. However, these feeds usually contain only the most recent items (e.g. the most recent 100 items added to the system). In other words, it is impossible to obtain, for example, a full bookmarking history of a particular document or a large number of documents that are assigned a particular tag. While Delicious also offers an official API that allows one to programmatically access its data via HTTP, this API only provides limited access to the data.[3]

In view of the above limitations, we collect data from Delicious by building a crawler program developed in the Python programming language to collect data directly from the Web site of Delicious. We collect data in a tag-by-tag manner. This means that the crawler is designed to collect as many bookmarks as possible that have been assigned a particular tag each time when it is deployed, i.e. we use

[3]See `http://delicious.com/help/tools`.

```
<documents>
  <document url="http://www.foodandwine.com/articles/10-best-online-wine-shops"
    title="10 Best Online Wine Shops" users="39" tags="32">
    <bookmark user="rodgersp" tags="wine" date="2009-01-22" />
    <bookmark user="frubilicious" tags="" date="2008-12-14" />
    <bookmark user="jcao" tags="life, wine" date="2008-12-06" />
    <bookmark user="cyberjuris" tags="wine, web, online" date="2008-09-07" />
    ...
  </document>
  ...
</documents>
```

TABLE 3.1: An excerpt of the dataset corresponding to the tag `wine`. It shows 4 of the 39 users who have bookmarked the Web site *10 Best Online Wine Shops*, along with the tags they used. Note that a user can assign no tags, or one or more tags to a bookmark.

a particular tag as a 'seed'. The data collected for a particular bookmark involve the URL, the title, the descriptions and tags contributed by the user, as well as the time at which the bookmark was created. Delicious itself advices users that requests to the server should be separated by an interval of at least one second to avoid being throttled (i.e. denied access to the server). However, in practice it may require up to six or seven seconds between requests to avoid denial of access in the middle of a long crawl. Therefore it might take up to several hours or a day to collect the data of a particular tag. Our major data sets were collected in the period between January and March 2009.

We store the data in text files in XML format with a schema designed to suit the purpose, with reference to the format described by Michael Noll who has developed an unofficial API to data in Delicious.[4] Table 3.1 shows an excerpt of the dataset corresponding to the tag `wine`. It should be noted that while it is possible to obtain the time at which a bookmark is created by retrieving an RSS feed, crawling the Web site of Delicious only returns the date on which the bookmark is created. The effects of this limitation on one of our studies will be further discussed in Chapter 5.

We first monitor the front page of Delicious as well as the page that lists a set of the most popular tags in order to collect a primary set of tags. This primary set of tags consists of several hundred tags. We then randomly select a total of 100 English tags from the set (some popular tags are words in other languages). In addition, we also pick another 35 tags that are observed to have multiple meanings in Delicious for the study of the social meanings of tags as described in Chapter 4. We thus have a total of 135 'seed' tags for our data collection task (see Table 3.2). We then deploy our crawler for each of the tags and collect a total of 135 data sets

[4]Delicious Python API: `http://www.michael-noll.com/wiki/Del.icio.us_Python_API`

.net, 3d, admin, adobe, advertising, **ajax**, algorithms, api, **apple**, **architecture**, argument, art, articles, **asl**, audio, **bath**, blogs, books, **bridge**, browser, business, **cambridge**, car, **chinese**, citation, climate, cms, **collection**, comics, computer, **convention**, cooking, cool, culture, ecommerce, economics, electronics, **english**, email, entertainment, environment, fashion, film, finance, firebug, firefox, flash, flickr, food, **forum**, **framework**, **free**, freelance, freeware, fun, funny, **gallery**, **games**, geek, geography, google, government, graphics, graphs, **green**, **guide**, hal, hardware, health, history, home, hosting, house, howto, html, humor, icons, ie, illustration, illustrator, im, images, information, inspiration, interactive, interesting, internet, iphone, **japan**, java, javascript, jobs, jquery, jvm, kernel, kids, **language**, later, learning, library, list, materials, media, mention, money, nu, online, opensource, **opera**, **phoenix**, photography, **player**, politics, programming, **race**, **recovery**, **rest**, **rice**, semanticweb, **sf**, **soap**, **streaming**, **sun**, svn, todo, travel, **tube**, tutorial, tv, ubuntu, **underground**, web, webdesign, **wine**, **xp**

TABLE 3.2: The 135 seed tags used in the data collection process. The bold tags refer to the 35 hand-picked tags.

for the experiments and analysis described in this thesis. Altogether, the data sets involve over 1,000,000 unique users, over 100,000 unique URLs (documents), and over 800,000 unique tags.

One of the challenges of analysing datasets collected from Delicious is that Delicious is growing fast and new bookmarks are continuously added to the system. This means that, as Heymann et al. (2008a) point out, without accessing to the complete set of data of Delicious any snapshot of the data is very likely to be subjected to imprecision or biases to certain extent. For example, by crawling data from the Delicious homepage the collected data will very likely be biased to active users and popular tags and URLs in the system, as Delicious tends to present popular URLs on its homepage. In the context of this thesis, this kind of bias may come from the selection of the tags we want to analyse. Instead of aiming at understanding the general characteristics of folksonomies, which as we have reported in Chapter 2 quite a number of studies have covered, we study folksonomies at the individual tag level, meaning that we focus on one tag at a time. Judging from our list of 'seed' tags, our data sets do cover a very broad range of topics. While there are tags about computer-related technologies (e.g. `java` and `semanticweb`), there are also tags about things in daily life (e.g. `health` and `tv`). Hence, we believe our analysis and experiments based on these data sets will return qualitative conclusions that can be applied to collaborative tagging systems in general.

3.4 Chapter Summary

In this chapter we have introduced Delicious, one of the earliest and most popular collaborative tagging systems on the Web nowadays, and have discussed its characteristics and functionalities. We have also described in detail how we collect data from Delicious for the experiments and analyses carried out in this project, as well as the data sets we have collected. In the next three chapters, we will describe the different studies that make use of the data sets mentioned in this chapter, starting with the social meanings of tags.

Chapter 4

Social Meanings of Tags

Tags can be considered as the most important elements in collaborative tagging systems. Tags allow users to annotate and classify Web resources as they like. Without tags, these systems are only online stores of resources without any organisation. However, as tags are freely chosen by users instead of being selected from a controlled vocabulary, there is no guarantee that every tag corresponds to a single well-defined meaning. The users do not even have to adhere to the conventional meanings of the words. In other words, the meanings of the tags used in a collaborative tagging system are highly dependent on the users of the system. This is also the reason of why there are so many ambiguous tags—tags that are used to represent different things depending on the contexts in which they are used—in folksonomies. In order to facilitate better organisation and retrieval of tagged resources, it becomes necessary to understand the ***social meanings*** of the tags, i.e. the meanings of the tags intended by the community in which they are used. We use ***tag contextualisation*** to refer to the process of identifying the social meanings of tags in a folksonomy.

In this chapter, we first discuss how semantics of tags can be defined based on their associations with other tags, users and documents in a folksonomy. We then discuss the nature and different types of ambiguity of tags. We describe a preliminary study on two popular but ambiguous tags, namely `sf` and `wine` in Delicious. By studying the collective behaviour of the users who have used these two tags, we gain insight into how tags can be contextualised. Based on the results, we investigate how tags can be contextualised in an unsupervised manner by performing clustering on different network models of a folksonomy. In particular, we want to test the following hypothesis regarding tag meanings in collaborative tagging:

> **Hypothesis 1 (tag meaning)**: When modelling a folksonomy, networks that explicitly take the users' collective behaviour into account are better in capturing meaningful associations between different entities in a collaborative tagging system, and clustering analysis of these networks produces more accurate results regarding the meanings of tags intended by the users.

4.1 Semantics of Tags

The semantics of a word refers to the meaning of the word or how the word is interpreted (Saeed, 2003). It is very common that a single word can be used to refer to different things. In linguistic terms they are said to have a large semantic field. In order to correctly understand the meaning of such ambiguous words, it usually requires a proper context in which there is information that helps one to interpret the words.

As we have discussed in Chapter 2, tags contributed by users in a collaborative tagging system also exhibit ambiguity. A tag can be used by different users (or sometimes even the same user) to refer to different things. However, such ambiguity is of greater extent than that one would expect in common usage. Pinker (2008) mentions that words have meanings that are commonly agreed on. When asked to write in one's own words, it does not mean that one can use any words to express an idea without paying attention to what the words are actually intended to mean, but it means that one has the freedom to combine known words in one's own way. However, in collaborative tagging this is not always true. Tagging is as much a personal activity as a social and collaborative activity (Sen et al., 2006). A user can use a tag to refer to something that is completely unexpected given the common understanding of the word, or he/she can create a tag that is a completely new word and interpret this tag in a personal way. And of course, a user is not obliged to give definitions to the tags he/she creates for use in the collaborative tagging system (which would probably greatly reduce the appeal of collaborative tagging).

Dictionaries (and also lexicons and thesauri) can be limited when they are used to understanding the semantics of tags for similar reasons. Firstly, tags are not necessarily made up of words that exist in a dictionary when users are free to create and use new terms or phrases as they see fit. Secondly, even for existing words, the intended meaning of a word may not be available in any dictionaries For

example, the tag `tube` has been used by users in Delicious to describe Web sites that present streaming video clips uploaded by Web users, a usage made popular by the video sharing site YouTube. Most importantly, if we are to figure out what exactly the tags are intended to mean within a collaborative tagging system, we must study the tags by putting them back into the context in which they are used. This context consists of the documents to which the tags are assigned, the users who use these tags, and the other tags that have been used together with this tag. Hence, ultimately the associations between these three types of entities, and the derived associations between entities of the same type, shall tell us what the tags are intended to mean.

4.2 Word Associations and Folksonomies

A first idea of understanding the semantics of a tag using associations in a folksonomy would be to examine the co-occurrence relations between the tag and other tags. In other words, we can study how frequently a tag is used by the same user or on the same document with other tags, such kind of co-occurrence of words is referred to as first order co-occurrences. For example, when the tag `java` co-occurs with the tags `programming` and `software`, it becomes obvious that the tag is referring to the Java programming language. In fact, word associations have been extensively used in word sense disambiguation to identify the sense of a particular occurrence of a word in a sentence (Ide and Veronis, 1998). Words that appear in the vicinity of the word to be disambiguated constitute the context within which the meaning of the word can be interpreted correctly. Heylighen (2001) refers to meaning of a word defined by its associations with other words as the connotation of the word.

There are different methods of acquiring information about word associations, and these can be broadly divided into two classes: (1) free association and (2) co-occurrence analysis. The first class of methods, free association, tries to solicit word association information from human subjects. Mika (2007) mentions the Edinburgh Associative Thesaurus (EAT) (Kiss et al., 1973) as a similar construct of the association network of tags resulted from users using pairs of tags together on some documents.[1] The EAT is constructed by asking people to generate words that they immediately think of when presented with a stimulus, which is also a word. This generates a large network of words that encodes their empirical

[1]Edinburgh Associative Thesaurus (EAT): `http://www.eat.rl.ac.uk/`

associations. Such approach has been extensively used in psycholinguistic research to study the language facilities of the human mind. Other similar studies, which aim at determining the associations between common English words, include those by PAlermo and Jenkins (1964) and Nelson et al. (1998). In addition, the Web site wordassociation.org tries to solicit word associations from general Web users by asking them to type in a word they first think of when a random word is presented to them when they visit the site.[2] The Web site currently reports a vocabulary size of over 65,000 words with over 11,000,000 associations, although the data is not publicly available.

On the other hand, co-occurrence analysis aims at estimating associations between words by processing a large amount of textual content using a computer. Strength of associations between words can be estimated by how frequently two words appear together within a document or a sentence. For example, Church and Hanks (1990) propose to use the concept of mutual information to estimate word associations from computer readable corpora. In addition, latent semantic analysis (LSA) (Dumais et al., 1988) is a widely used mathematical method for modelling associations between words and sentences. LSA measures not only first order co-occurrences but also second and higher order co-occurrences (co-occurrence of co-occurred words) in order to find out the degrees of similarity between words appearing in a corpus.

Free association represents an effort to solicit associations between words from human subjects, while co-occurrence analysis aims at discovering this kind of information by automated processing of large corpora. Generating word associations from folksonomies can be considered as a combination of the above two approaches. The co-occurrences of tags in a folksonomy depend both on the users' choices as well as the content of documents being tagged. While users have the freedom to use any combination of tags together when performing tagging, their choices are somehow guided by the content of the documents. For example, while `tube` is highly associated with `pipe` in EAT, these two words will not be used together as tags to describe any documents if no documents address relevant topics.

While there are studies that perform large scale clustering on folksonomies to find out clusters of tags that address similar topics, very few studies so far have been carried out to understand the relationship between associations between tags and the multiple meanings of ambiguous tags. For example, it is straight forward to find out that the tag `sf` is highly associated with `sanfrancisco`, `bayarea`,

[2] `http://www.wordassociation.org/`

`sciencefiction` and `reading`. However, it is obvious that these associations must be interpreted within a suitable context. It is not clear how tags can be contextualised in a way that one can reveal the highly associated tags in different contexts (which probably correspond to the different meanings of an ambiguous tag).

In addition, co-occurrences between tags can be obtained by either examining how often two tags are used together by the same user (regardless of the documents on which they are used) or how often they are used together on the same document (regardless of the users who use them). This difference between user-based and document-based co-occurrences, however, is usually overlooked in the literature. Nevertheless, this is actually an important question when we try to understand the semantics of tags through their associations. Mika (2007) reveals that the tag network generated by user-based co-occurrence and that generated by document-based co-occurrence can be very different. Nevertheless, the effect of such difference on the task of revealing multiple meanings of tags is yet to be investigated.

To have a better understanding of the associations between tags, we argue that we have to pay attention to the contexts in which they are used, these include both the users and the documents that the tags are associated with. In fact, a folksonomy is a structure in which each of the three types of entities helps constitute a context in which the semantics of the others can be understood.

4.3 Tag Ambiguity

As we have discussed in Section 2.3.2.4, there are a lot of different types of tags being used in existing collaborative tagging systems like Delicious and Flickr. In addition to the fact that tags can be classified into different types, they also exhibit different forms of ambiguity.

Tags that have a number of different meanings when presented alone are very commonly in collaborative tagging systems. While this is expected as tags are after all made up of words that inherently usually possess more than one specific meaning, ambiguity in tags arises also because of some other reasons. As we have mentioned earlier in this chapter, tagging can be a very personal activity such that there can be idiosyncratic interpretations of common words. In addition, the user community may come up with new meanings for existing words in order to

better describe their favourite resources. Some authors use the concept of entropy (Shannon, 1949) in information theory to measure the extent to which a tag is ambiguous. However, such measure does not distinguish between different types of ambiguity of tags, which we would like to discuss in more details.

One of the most obvious forms of ambiguity in tags is that things interested by the users incidentally share the same abbreviation or acronym. For example, in Delicious we can see that the tag `sf` is used as an acronym for both 'San Francisco' and 'science fiction', and the tag `xp` is used to represent 'Windows XP' as well as 'extreme programming', and the tag `asl` can mean 'American sign language' or 'advanced squad leader' (a kind of war game).

Another common form of ambiguity is ambiguity of name entities. Names of people, places, organisations and companies may share the same form. For example, pages that are assigned the tag `cambridge` refer to the Cambridge in the United Kingdom or the Cambridge in Massachusetts in the United States. The tag `london` are used on pages about London, the capital of the United Kingdom, London, the city in Canada, or Jack London, the famous American writer. The tags `apple` and `sun` are representatives of cases in which a common noun is used as the name of a company, not to mention the well known example of `jaguar` that has appeared in many studies of word sense disambiguation and Web search result classification in the literature.

In addition, there are also a lot of other tags that possess different meanings, and they can generally be considered as polysemes or homonyms. For example, the tag `tube` can be observed to be used on pages about the London Underground or on Web sites for video clips sharing, and the tag `bridge` is used to refer to several concepts, including physical architectural structures, a kind of card game, and a kind of design pattern in software engineering.

Finally, there are also many tags that correspond to some very general or broad concepts such that they have many sub-topics that can be explored. Strictly speaking, these tags are not ambiguous because they are always used to refer to the same concepts. However, since these tags refer to very broad categories, the topics addressed by documents assigned these tags are very diverse. For example, the tag `books` is used in Delicious on documents about science fictions, classic novels, cooking recipes, and reference books of programming languages to name a few. In addition, some tags such as `todo` and `toread` that are used for organisational purposes can be assigned to documents of any topics.

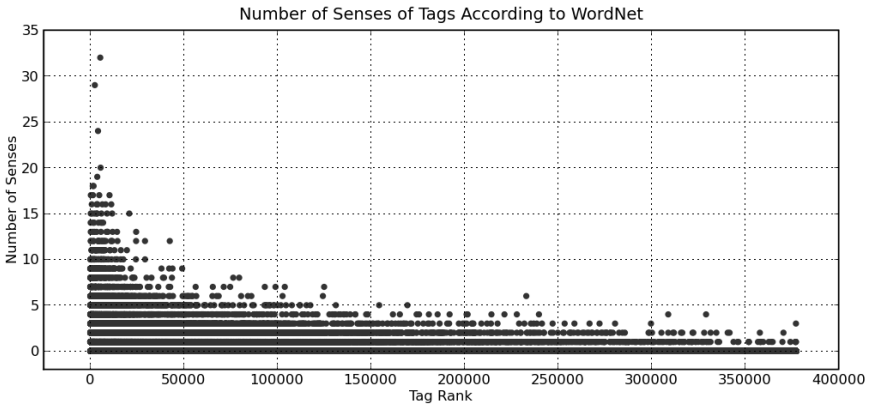


FIGURE 4.1: Number of senses of tags in Delicious according to WordNet. The x-axis refers to the ranks of the tags according to their occurrences. Rank 0 refers to the highest rank, and a tag is less common as this value increases.

To understand how ambiguous tags can be, we conduct an experiment to look into the ambiguity of tags in Delicious with the help of WordNet (Miller, 1995), an English lexical database that can be queried for the different senses of a given word.

We develop a crawler program using Python and use it to access pages in Delicious and collect a total of 809,116 tags that have been used by more than two users in Delicious (otherwise Delicious would not show how many times the tag has been used). We filter this set of tags and extract tags that are only composed of English alphabets, resulting in a refined set of 377,344 tags. For each of these tags, we access its corresponding page in Delicious and get the number of times it has been used in Delicious. We then query WordNet using these tags and obtain the number of senses they have.

Figure 4.1 shows a scatter plot of the number of senses of a tag against its rank. We have to mention that out of the 377,344 tags only 19,923 of them are found to appear in WordNet. However, the scatter plot shows a clear pattern that tags that are more popular tend to have more senses than less popular tags, suggesting that popular tags can be more ambiguous. The two variables show a slight negative correlation of $r = -0.244$. There are several possible reasons for such correlation. Firstly, a more ambiguous tag (having more different senses) can be used in more different contexts and is therefore more likely to be used by more users in Delicious. Secondly, as we have mentioned in Chapter 2, some authors have reported that popular tags mostly represent broad categories. Names of broad categories are more general and are therefore more likely to have a greater number of senses.

The above experiment, however, provides only a rough estimation of the extent of ambiguity of tags in Delicious for two reasons. Firstly, it does not tell us how ambiguous tags that do not appear in WordNet are. Secondly, a tag having a lot of senses in WordNet does not necessarily mean that it is ambiguous in Delicious. The ambiguity of a tag within a collaborative tagging system actually depends on the resources and the community of users of the system. A tag possessing different meanings may not be ambiguous if all or the vast majority of users in the system only use it to refer to one single concept. For example, while the tag `ajax` represents the name of a football team as well as that of a Web technology, the vast majority of users in Delicious only refer to the latter sense of the tag. Nevertheless, this experiment does tell us something about the ambiguity of tags: it affects mainly popular tags and is therefore an important issue to most of the users of a collaborative tagging system. In addition, it highlights an important thing that we should pay attention to when trying to understand the meanings of a tag, which is that even such a rather comprehensive dictionary as WordNet can only provide the meanings of about 5% of the tags, due to the large number of user-invented tags such as tags that are made up of new terms or multiple words. Even considering the possible increase in this proportion tag cleaning and stemming may lead to, external resources still seem to be very inadequate in this respect.

4.4 Networks in Folksonomies

To investigate the associations between different entities in a folksonomy, it is best to represent a folksonomy as a graph or a network, with vertices representing the entities and edges representing their associations. Recall that a folksonomy is defined as a tuple $\mathcal{F} = (U, D, T, R)$, where U is a set of users, D is a set of documents, T is a set of tags, and $R \subseteq U \times D \times T$ is a set of annotations representing a user's tagging a document with a particular tag. As folksonomies involves mainly three different types of nodes, their underlying networks are usually in the form of tripartite hypergraphs (Catutto et al., 2007; Lambiotte and Ausloos, 2006; Mika, 2007; Niwa et al., 2007): $\mathcal{H} = \langle V, E \rangle$ where $V = (U \cup T \cup D)$ and $E = \{(u, t, d) | (u, t, d) \in R\}$.

Depending on the types of entities in a folksonomy one would like to focus on, different types of sub-networks can be generated from the tripartite hypergraph of a folksonomy. For example, a network of tags with weights of edges determined by

co-occurrence is considered in quite a number of studies in the literature (Begelman et al., 2006; Heymann and Garcia-Molina, 2006; Shen and Wu, 2005). On the other hand, Mika (2007) considers the bipartite graphs of user-tag associations and document-tag associations, which are further folded into a one-mode network of tags as a lightweight ontology. Here, as we are focusing on individual tags in order to uncover their multiple meanings, we will always be working with a subset of the folksonomy which is associated with a particular tag. A network can always be represented by an adjacency matrix. We will discuss the matrix representation of a network later in this chapter when it is needed.

Before going into the details of the experiments and analysis, we introduce some notations that will be used in the rest of this chapter. Given a tag t, we denote by U_t the set of users who have used the tag t on one or more documents:

$$U_t = \{u | \exists d \in D, (u, t, d) \in R\} \tag{4.1}$$

by D_t the set of documents which have been assigned the tag t:

$$D_t = \{d | \exists u \in U, (u, t, d) \in R\} \tag{4.2}$$

and by T_t the set of tags which have been used together with t on some documents by the same users:

$$T_t = \{t' | \exists (u, d) \in U \times D, (u, t, d) \in R \wedge (u, t', d) \in R \wedge t \neq t'\} \tag{4.3}$$

4.5 Preliminary Studies

To investigate how tags can be contextualised in a folksonomy by exploiting the collective behaviour of the users, we perform a preliminary study on two tags, namely `sf` and `wine`. In this study, we examine the networks of users and documents associated with these two tags, and attempt to understand how different concepts associated with the tag can be discovered.

The reasons of choosing the tags `sf` and `wine` are twofold. Firstly, they are both very popular tags in Delicious, with each of them having been used for over 1 million times. Secondly, both tags are observed to be ambiguous as they are used to represent multiple concepts in Delicious: `sf` has been observed to be used to represent 'science fiction' and 'San Francisco', while `wine` has been observed to be

used to represent both a kind of alcoholic drink and a Linux software package.[3] We expect these two features of the tags will lead to clearer results for further analysis.

In this exploratory study, we want to gain a better understanding on whether users are likely to use a tag in a consistent way, i.e. they will use a tag to represent the same thing for the most of the times. Users use tags to describe their favourite resources and therefore it makes little sense to use a tag to refer to multiple concepts as this will only complicate the process of retrieval. The aim is thus to investigate to what extent this assumption about the behaviour of the users is correct. The result will form the basis of the experiments on tag contextualisation that will be described later in this chapter.

4.5.1 Document and User Networks

We single out the two data sets of `sf` and `wine` from the 135 data sets we have collected from Delicious. The data set of `sf` involves a total of 64,185 users, 1,530 documents, and 19,587 tags. On the other hand, the data set of `wine` involves a total of 41,742 users, 1,481 documents, and 9,856 tags.

We first construct a network representation of the documents in the data sets. In this network, two documents can be considered as related (connected) to each other if there is a user who has assigned the tag in question (which is the tag `sf` or `wine` in this study) to both of the two documents. The edge between the two vertices representing these two documents can be weighted by the number of such users. In mathematical notations, such a network of documents can be represented by a matrix $\mathbf{B} = \{b_{ij}\}$, where

$$b_{ij} = |\{u|(u, t, d_i) \in A \wedge (u, t, d_j) \in R\}|. \tag{4.4}$$

On the other hand, we can also consider a network of users in the data sets. In such a network, two users can be considered as connected to each other if they both use the tag in question on the same document. The edge between the two users can be weighted by the number of such documents. Such a network can be considered as an implicit social network of the users with respect to a particular tag. They are connected because it is very probable that they share the same interpretation of the tag in question, because they both use the tag on the same documents,

[3]The Web site of Wine, the software package, can be found at `http://www.winehq.org/`.

suggesting that the tag means the same thing to both of them. In mathematical notations, this implicit social network can be represented by a matrix $\mathbf{S} = \{s_{ij}\}$, where

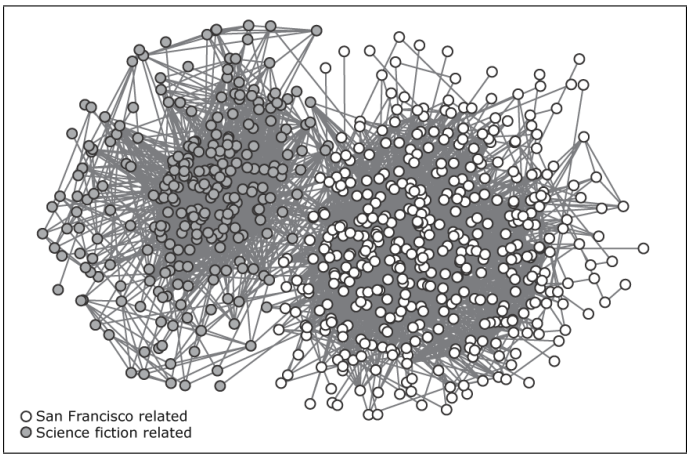
$$s_{ij} = |\{d|(u_i, t, d) \in A \wedge (u_j, t, d) \in R\}|. \tag{4.5}$$

Firstly, we generate the networks of documents for each of the tags `sf` and `wine`, and feed the data into the network analysis package Pajek (de Nooy et al., 2005).[4] We visualise the networks using the Kamada-Kawai layout algorithm (Kamada and Kawai, 1989), which puts highly connected vertices closer to each other based on energy minimisation, thus producing a clear picture of the networks for inspection of possible clusters of vertices. The resultant networks consist of a lot of disconnected components, most of which are isolated documents that are not associated with the others. Figure 4.2(a) shows the largest component in the document network for `sf`, and Figure 4.2(b) shows the largest component in the document networks for `wine`.

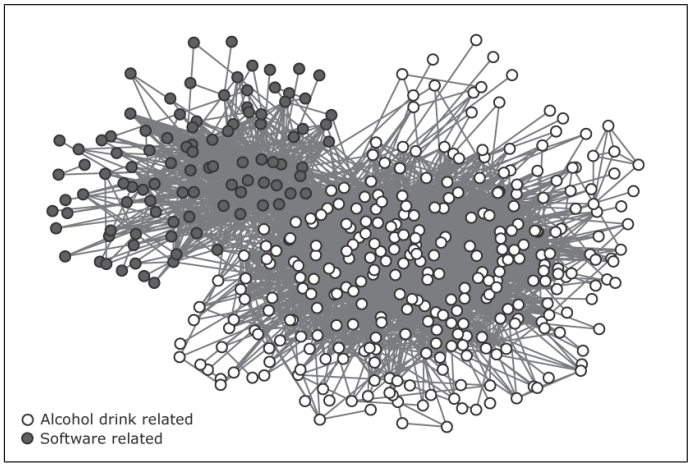
We first take a look at the network of documents for the tag `sf`. Two large clusters of vertices can be observed in their document networks. One hypothesis for the existence of clusters in the network of documents is that they correspond to groups of documents related to the different concepts the tag *sf* is used to represent. A similar hypothesis can be applied to the network of users: the different clusters correspond to different groups of users who have used the tag `sf` to represent different concepts.

Since documents are connected if a user has assigned the tag `sf` to them, this implies that connected documents are considered by the user to be related to the same concept represented by `sf`. In addition, if we assume that a user would be consistent in using the same tag for the same concept, it is reasonable to suggest that documents within the same cluster would address the same topic represented by the tag `sf`. As we understand through observation that two major concepts, namely 'science fiction' and 'San Francisco', are associated with `sf`, we can further suggest that the two major clusters in the network correspond to documents related to 'science fiction' and 'San Francisco' respectively. Similarly, the two clusters observed in the network of documents for the tag `wine` can also be hypothesised to be corresponding to the two things represented by the tag in Delicious, namely (1) alcoholic drink and (2) a software application. To verify this hypothesis, we perform further analysis on the data.

[4]Pajek is available from `http://vlado.fmf.uni-lj.si/pub/networks/pajek/`



(a) The largest component in the network of **documents** of the tag `sf` (a total of 714 documents). White vertices represent documents related to San Francisco, while gray vertices represent documents related to science fiction.

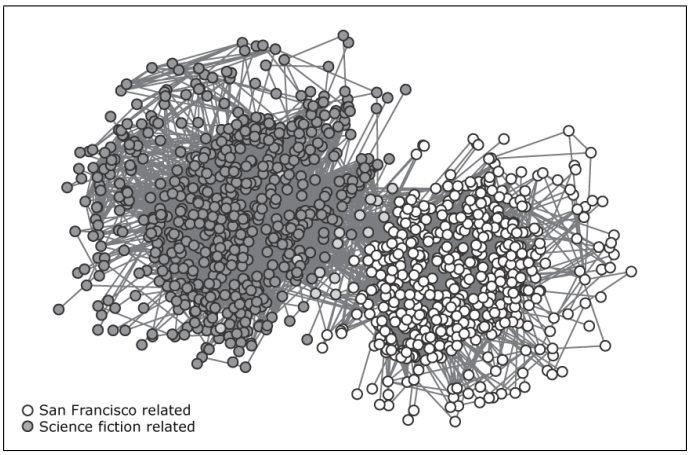


(b) The largest component in the network of **documents** of the tag `wine` (a total of 504 documents). White vertices represent documents related to alcoholic drinks, while gray vertices represent documents related to the software application named Wine.

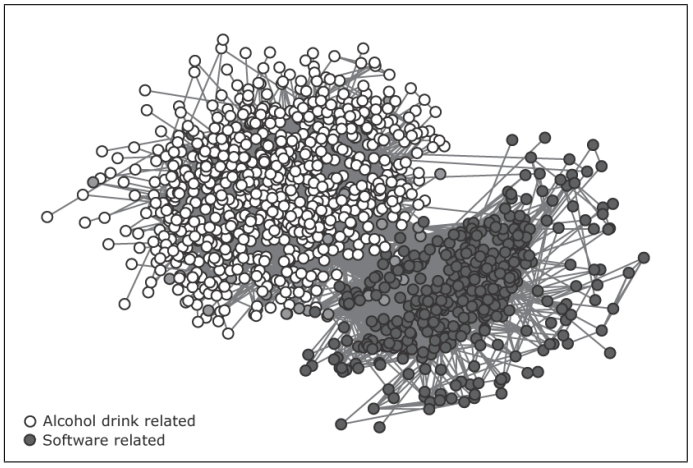
FIGURE 4.2: Networks of documents for the tags `sf` and `wine`.

Firstly, we manually examine all the Web documents represented by the vertices in the two networks of documents. We classify the documents based on their content as well as the tags assigned to them by the users. After that, we combine the information with the original networks, and use different colours of the vertices to indicate the different topics addressed by the documents, as shown in Figure 4.2. The result shows that documents that address the same topic are indeed grouped into the same cluster.

It is interesting to note that there are actually some edges running between the two clusters. These connections show that some users do use the same tag to describe documents that address different topics represented by the tag, implying that these users use the tag in an ambiguous way. To gain a more thorough understanding of



(a) The largest component in the network of **users** of the tag `sf` (a total of 4,175 users). White vertices represent users who have assigned `sf` only to documents related to San Francisco, while dark gray vertices represent users who have assigned the tag only to documents related to science fiction. Light gray vertices represent users assigning the tag to both types of documents.



(b) The largest component in the network of **users** of the tag `wine` (a total of 5,321 users). White vertices represent users who have assigned `wine` only to documents related to alcoholic drinks, while dark gray vertices represent users who have assigned the tag only to documents related to the software application named Wine. Light gray vertices represent users assigning the tag to both types of documents.

FIGURE 4.3: Networks of users for the tag `sf` and `wine`.

how users use these two tags, we identify the users who have assigned the tags to the documents represented by the vertices in the document networks, and check how many users have actually assigned the tags to different groups of documents. We also generate the networks of users for the two tags and use different colours to indicate whether they are consistent in using the tags (see Figure 4.3).

We find that, among the 8,267 users who have assigned the tag `sf` to the documents in the network shown in Figure 4.3(a), 3,309 users have only assigned the tag to documents about science fictions, while 4,931 users have only assigned the tag to documents about San Francisco. There is only 27 users who have assigned the

tag to both documents about science fiction and those about San Francisco. The analysis of the users for the tag `wine` returns similar results. Among the 11,346 users who have assigned the tag `wine` to the documents in the network shown in Figure 4.3(b), 6,551 users have used the tag to refer to only alcoholic drinks, while 4,671 users have used the tag to refer to only the software application named Wine. There is only a total of 124 users who have used the tag to refer to both things. These results show that despite the fact that the different contexts in which these tags are used are all quite popular among the users, the users have been quite consistent in using these tags.

4.5.2 Discussion

The above analysis of the documents and users associated with the tags `sf` and `wine` reveals several interesting aspects of Delicious. Firstly, it shows that, by identifying different clusters of documents or users in a network representation of a subset of a folksonomy, we are able to discover the different meanings of an ambiguous tag. It is particularly important to note that in the above analysis no information about the content of the documents or other tags that have been assigned to the documents is considered in the process. The analysis only involves characterising documents by the users who are interested in them and have assigned the tag in question to them. This reveals that even by looking at the collective user behaviour alone we get to know a lot about the semantics of a tag.

The preliminary study reveals that few users would use a tag in an ambiguous way, i.e. using it to refer to one thing at one time and another thing at another time. There can be several reasons to this. Firstly, the users may be aware of the problems associated with using the tag ambiguously (retrieval of relevant documents using the tag becomes harder), such that they tend to use the tag in a consistent way. Secondly, the majority of users in our data sets may only be interested in one of the concepts represented by the tags. For example, a user interested in Web pages presenting information about different kinds of red or white wine may not be interested in, or even know about, the software application named Wine, and therefore the idea that the tag `wine` can also mean something else would not come to his/her mind.

No matter whether users are consciously consistent in using a tag or are only interested in one particular meaning of a tag, given enough number of users using an ambiguous tag, we see that clusters of documents and users corresponding to

the different meanings of the tag would emerge. In fact, we can see collaborative tagging as a process through which users collectively come to define a tag by assigning the tag, as well as other contextually related tags, to different documents, giving rise to co-occurrence patterns that reveal the different meanings of the tag. In the following sections, we describe an experiment of larger scale that aims at investigating which types of network representation of a folksonomy are best at revealing this kind of social meanings of ambiguous tags.

4.6 Tag Contextualisation

Our preliminary study suggests that identifying the different contexts in which a tag is used by the users can be done by performing clustering on the networks induced from folksonomies. However, constructing a network of documents based on the users is only one of the many ways of representing a folksonomy as a network. In fact, tag contextualisation—the task of identifying the different contexts in which a tag is used—can be done using many different ways, such as by studying networks of users, tags or documents. There have been no studies that compare the characteristics of these different networks and their usefulness in revealing the multiple meanings of ambiguous tags. Usually a particular network is chosen in an ad-hoc fashion, and sometimes the ways these networks are constructed are not paid much attention. Nevertheless, we believe that the differences between these networks do affect the outcomes of tag contextualisation.

In fact, it is pointed out that networks of tags that explicitly take the social context into consideration are better in revealing the semantic relations between the tags (Mika, 2007). In the following experiments we take this notion further and study whether this is true in understanding the semantics of individual tags. We consider several different types of networks induced from a folksonomy and compare their performance in the task of tag contextualisation. Through the following experiment, we also want to investigate whether it is feasible to identify the multiple meanings of ambiguous tags using an unsupervised method, instead of consulting external resources, to avoid the drawbacks described earlier in this chapter.

4.6.1 Network Models of Folksonomies

The aim of representing a (subset of) folksonomy as a network is to reveal the associations between different entities in the folksonomy. These associations come in different kinds. For example, there are direct associations between users and documents as they assign tags to the documents, or between users and tags as the users use the tags on some documents. There are also indirect or implicit associations between tags, as they are used together on some documents by some users. Tags can also be associated with each other if their co-occurring tags are very similar. All these different considerations give rise to different kinds of networks, which are all likely to reveal the clustering structures in a folksonomy and are therefore useful for tag contextualisation. In the following, we describe several different kinds of networks, describe how they can be constructed and explain their implications for the task of tag contextualisation.

4.6.1.1 Tag-based Document Networks

As we have discussed in Chapter 2 Section 2.1, tagging can be considered as an act of indexing the documents on the Web, and therefore tags can be used to characterise documents in the same way as keywords are used to characterise documents in information retrieval tasks. A weighted term (tag) vector $\boldsymbol{v}_d$, which is commonly used in document clustering and information retrieval (Cutting et al., 1992; Stefanowski and Weiss, 2003; van Rijsbergen, 1979), can be constructed to represent a document d, with each element of the vector corresponding to the number of times a tag has been assigned to it.

$$\boldsymbol{v}_d = (v_{d,1}, v_{d,2}, ..., v_{d,|T_t|}) \tag{4.6}$$

where $v_{d,i} = |\{u|(u, t_i, d) \in R\}|$.

A similarity matrix $\mathbf{A} = \{a_{ij}\}$ can be constructed to represent the pairwise similarity of each document by using the cosine similarity measure:

$$a_{ij} = csim(\boldsymbol{v}_{d_i}, \boldsymbol{v}_{d_j}). \tag{4.7}$$

where

$$csim(\boldsymbol{v}_1, \boldsymbol{v}_2) = \frac{\boldsymbol{v}_1 \cdot \boldsymbol{v}_2}{||\boldsymbol{v}_1|| \times ||\boldsymbol{v}_2||} \tag{4.8}$$

and $\boldsymbol{v}_1, \boldsymbol{v}_2 \in \mathbf{R}^n$.

In this way, a network with $|D_t|$ vertices representing each of the documents and edges weighted by the similarity between these documents can be constructed. It can be hypothesised that, if the tag t is used by the users to refer to different concepts in different contexts, we should find vertices representing documents which correspond to the same context to be highly connected with each other, resulting in different clusters of vertices. To obtain a label for each of the clusters, one can extract the tags which are most frequently used among the documents within the clusters.

It should be noted that the tag t which we are looking at is not included in the term vectors given the definition of T_t. This is actually desirable because t is assigned to every document (and probably by many users) in the set D_t, therefore the inclusion of the tag in the vectors will probably result in all the documents being very similar to each other.

In addition, when constructing the term vectors we only consider the frequencies of the tags, while some weighting schemes such as the TF-IDF (term frequency-inverse document frequency) scheme (Salton and Buckley, 1987) can also be used, as demonstrated, for example, by Brooks and Montanez (2006) and Cattuto et al. (2007). Our reason is that we are trying to group documents which are about similar topics (e.g. San Francisco), instead of trying to identify keywords which are most important to a document. It is found that tags are more likely to be broad terms rather than specific terms (Noll and Meinel, 2008a), suggesting that tags are more likely to be used to categorise a document. Hence, by considering the frequencies of tags in term vectors, we should be able to group documents into different categories, which correspond to the different contexts in which the tag t is used.

4.6.1.2 User-based Document Networks

The second type of network is the document networks we have discussed in Section 4.5. These networks are constructed based on the consideration that documents in a tagging system can also be characterised by the users who have assigned tags to them, and that documents tagged by similar users can be considered as similar to each other. As a reminder, this kind of networks can be represented by a similarity matrix $\mathbf{B} = \{b_{ij}\}$:

$$b_{ij} = |\{u|(u,t,d_i) \in A \wedge (u,t,d_j) \in R\}|. \tag{4.9}$$

In this way, a network with $|D_t|$ vertices representing the documents and edges weighted by the number of users who have assigned the tag to both of them can be constructed.

While this type of network that characterises documents simply by the users who have assigned a particular tag to them has not been considered in any previous works, it does provide valuable insight into how the tag is used among the users by putting it into the social context. As our preliminary study on this type of network has shown, most users are consistent in using a certain tag, meaning that they are unlikely to use the same tag to refer to different concepts. Hence, documents that are linked to each other in this network are likely to be about the same topic and constitute the same context in which the tag is used.

If clustering algorithms applied to this type of network reveal any clusters of documents, it is very likely that they will correspond to different contexts in which the tag is used. To obtain a label for each cluster, we can extract the tags which are most frequently used among the documents within the clusters.

4.6.1.3 Tag Co-occurrence Networks

Besides document networks, we can also look directly at the set T_t of tags that are used together with the tag t. The most common way of constructing a network of tags that reflects the relations between them is to consider their co-occurrence (Begelman et al., 2006). Intuitively two tags are more related to each other if they are used together more frequently on the same documents and/or by the same users. Mathematically, a matrix $\mathbf{C} = \{c_{ij}\}$ representing the strength of relations between tags can be constructed by counting the number of times two tags are used together using one of the following two methods:

$$c_{ij} = |\{d|\exists u_a, u_b, (u_a, t_i, d) \in R \wedge (u_b, t_j, d) \in R\}| \quad (4.10)$$

$$c'_{ij} = |\{(u, d)|(u, t_i, d) \in R \wedge (u, t_j, d) \in R\}| \quad (4.11)$$

with the exception that $c_{ij} = 0$ if $i = j$.

While Equation 4.10 only requires two tags to be assigned to the same document for the situation to be considered a co-occurrence, Equation 4.11 defines co-occurrence between two tags as a situation in which they have to be used on the same document by the same user. This distinction is not explicitly considered and discussed in previous studies. A decision on which one of these two methods

are used is often made arbitrarily. For example, the former is used by Begelman et al. (2006) in tag clustering, and the latter is used by Cattuto et al. (2008a) when studying various tag similarity measures. However, we believe that there are differences in the resultant networks constructed by using these different methods.

In the experiment to be described latter in this chapter, we consider both methods and want to find out whether the associations established by the users are significant in understanding the semantics of a tag. Equation 4.11 will produce tag relations with smaller weight because obviously $c'_{ij} \leq c_{ij}$ (tags can be assigned to the same document but not necessarily by the same user). However, it can be hypothesised that Equation 4.11 will produce tag relations of higher significance because the viewpoints of the users are explicitly taken into account. In addition, Equation 4.11 should be less vulnerable to spamming in collaborative tagging systems, as tags assigned by spammers are much less likely to be associated with tags assigned by other users (Koutrika et al., 2008).

It should be noted that there is also one more type of tag co-occurrence, which is the situation in which two tags have been used by the same user, regardless of whether they have been used on the same document or on different documents. However, we believe that this kind of tag co-occurrence is of relatively less importance here. In a separate study (Au Yeung et al., 2009b), we reveal that user interests in collaborative tagging systems, and in particular in Delicious, can be very diverse. This is reflected in several experiments that measure the diversity and co-occurrence frequencies of the tags used by the users. In other words, tags used by the same user are very likely to be words taken from different domains. Hence, associating tags based only on the users who have used them is not likely to produce useful semantic relations between the tags.

4.6.1.4 Tag Context Similarity Networks

The last type of network we consider in this chapter is based on the distributional measure of tag relatedness described by Cattuto et al. (2008b). In order to use this measure, we have to define a tag co-occurrence vector $\boldsymbol{v}_{t_i}$ for each tag $t_i \in T_t$:

$$\boldsymbol{v}_{t_i} = (v_{t_i,1}, v_{t_i,2}, ..., v_{t_i,|T_t|}) \tag{4.12}$$

where $v_{t_i,j} = c_{ij}$ or $v_{t_i,j} = c'_{ij}$ depending on which of the aforementioned method is used. A matrix $\mathbf{D} = \{d_{ij}\}$ representing a network of tags can then be constructed

by calculating the similarity between two tags with the cosine similarity measure:

$$d_{ij} = csim(\boldsymbol{v}_{t_i}, \boldsymbol{v}_{t_j}) \tag{4.13}$$

The tag co-occurrence vector reflects the context in which a tag is used because it encodes the co-occurrence frequencies of other tags which are used with this tag. Hence, the cosine similarity used in Equation 4.13 is actually performing a comparison of the contexts in which two tags are used (Schütze, 1998). This is different from the tag co-occurrence network in which tags are considered to be related or similar simply when they are used together.

The networks we consider here are constructed based on the *fully-connected* approach (Luxburg, 2007), which means that any pair of vertices with a positive similarity value between them will be connected by an edge. In fact there are other ways to construct these networks. In particular, we can choose to discard certain edges if their weights are too small. For example, we can adopt the ϵ-neighbourhood approach, which removes edges with weights lower than ϵ. Or we can adopt the k-nearest neighbour approach, in which each vertex in the graph connects to at most k neighbours which are most similar to it. However, as it is not clear at present which approach is the most suitable for our tasks and we would like to take as much information as we have into consideration, the fully-connected approach is used.

4.6.2 Network Clustering

In the networks constructed using the methods described in the previous sections, vertices that correspond to the same meaning of the tag in question should be highly connected with each other, and vertices that correspond to different meanings of the same tag should only be loosely connected with each other. It is desirable to reveal these different clusters, or in other words the community structures in these networks, such that we can identify the different groups of documents/users/tags that correspond to the different meanings of a tag, and ultimately achieving our purpose of tag contextualisation.

The task of clustering vertices in a network is a well-researched area in different fields, including mathematics, physics and computer science, and is a special case of the more general task of cluster analysis in data mining. There are in fact a

lot of algorithms available for clustering vertices in a network (Newman et al., 2006; Radicchi et al., 2004). Some of these algorithms require the number of clusters to be achieved at the end of the process to be specified at the beginning (e.g. the k-means algorithm), while some require a certain threshold at which a tree-like structure of the vertices is cut to obtain different groups of data points (e.g. hierarchical clustering algorithms). There are also algorithms that are fully unsupervised in the sense that they can come up with an optimal number of clusters based on some measures of the 'goodness' of the final division of the data points.

For the task of tag contextualisation, it is quite impossible to specify the number of clusters at the very beginning. This is because we have little idea of the number of meanings of a particular tag used by the users in a collaborative tagging system. Hence, it is desirable that a clustering process would produce an optimal number of clusters that reveal the different meanings of a tag. Such requirement is very common in practical applications in which the number of clusters or communities of vertices are not known ahead of time (Newman and Girvan, 2004). In fact, the task of network clustering, or now more commonly known as community discovery in networks, has attracted much attention in recent years due to its wide application in such different areas as physics, biology, citation analysis, and computer science (Newman and Girvan, 2004; Radicchi et al., 2004). Many algorithms for revealing the community structures in networks are developed. In particular, a method called ***modularity*** that measures the 'goodness' of a division of the vertices in a network is widely studied. We describe research in this area in more detail in the following section, and go on to discuss how we apply a community discovery algorithm to our task of tag contextualisation.

4.6.2.1 Community Discovery Algorithms

Network structures can be found in many real life situations. Biological systems such as networks of molecular interactions (Holme et al., 2003), social networks (de Nooy et al., 2005) and the World Wide Web (Flake et al., 2002) are common examples. One common feature of all these different kinds of networks is the existence of community structures. Community structure refers to the characteristics that vertices within a network tend to come together to form groups, and connections between vertices within groups are denser than those between groups (Newman and Girvan, 2004). It is of interest to discover the communities within a network, because this usually allows the characteristics of the network and the behaviour

of the individual elements to be better understood. In particular, a community is likely to correspond to a group of vertices with similar features (Newman, 2006).

Approaches to identifying communities within a network can be divided into two main categories, namely agglomerative and divisive algorithms (Radicchi et al., 2004). Agglomerative algorithms such as hierarchical clustering (Berkhin, 2002) consider a network with isolated nodes in the beginning, and then iteratively add edges to the network to connect the nodes, starting from the nodes which are considered to be closest or most similar to each other. In this way, larger and larger groups of vertices are obtained. The result of such algorithm is usually represented in the form of a dendrogram. Communities can be obtained by cutting the dendrogram at an appropriate level. On the other hand, divisive algorithms work on the problem in the reverse direction. These algorithms start from the original network and iteratively remove edges connecting the nodes. Edges that are likely to connect nodes from different clusters are removed first (Newman and Girvan, 2004). In this way, the network is gradually divided into separate components, revealing the underlying community structure.

One crucial step in discovering the underlying community structure in a network by either agglomerative or divisive algorithms is the point at which the process should terminate. For an agglomerative algorithm, we have to determine at which level of the resultant dendrogram we should administer a cut in order to reveal the underlying community structure. This is also true for a divisive algorithm for if we cannot determine at which point we should stop removing edges we would end up removing all the edges in the network and achieve no meaningful results. To determine if a particular division of a network is the best, i.e. comes closest to reveal the underlying community structure of the network (if there is any), the measure of modularity (Newman and Girvan, 2004) is usually used.

Modularity offers a quantitative way to evaluate the 'goodness' of a certain division of a network. The basic idea of calculating modularity involves comparing the actual number of edges within a community with the expected number of edges if they are placed in a random manner. Since communities are groups of vertices which are more closely connected with each other than with vertices in different communities, the number of edges within a community should be higher than that in the case of randomly placed edges.

Here we present a formal definition of the measure of modularity (Newman and Girvan, 2004). Firstly, we define A as the adjacency matrix of a given network,

with a_{ij} as its elements.

$$a_{ij} = \begin{cases} 1 & \text{if vertices } i \text{ and } j \text{ are connected,} \\ 0 & \text{otherwise} \end{cases} \tag{4.14}$$

Secondly, we define k_i as the degree of vertex i.

$$k_i = \sum_j a_{ij} \tag{4.15}$$

Also, we define a function δ which tells us whether two vertices are placed under the same community given a particular division of the network.

$$\delta(c_i, c_j) = \begin{cases} 1 & \text{if } c_i = c_j \\ 0 & \text{if } c_i \neq c_j \end{cases} \tag{4.16}$$

where c_i and c_j refer to the clusters which vertices i and j belongs to respectively.

The modularity of a division of a network is then given by:

$$Q = \frac{1}{2m} \sum_{ij} \left(a_{ij} - \frac{k_i k_j}{2m} \right) \delta(c_i, c_j) \tag{4.17}$$

where

$$m = \frac{1}{2} \sum_{ij} a_{ij} \tag{4.18}$$

is the total number of edges in the network.

With modularity as a quantitative measure of the quality of a particular division, the problem of community discovery can then be considered as a problem of optimising the value of modularity over all possible division of a network. However, for many networks the scale is so large that it would be impossible to calculate the modularity of every possible division in order to find the best answer. In view of this problem, different heuristics have been proposed to optimise modularity in combination with either an agglomerative or a divisive algorithm. For example, Guimera and Amaral (2005) propose to use simulated annealing to optimise modularity. Newman and Girvan (2004) propose a divisive algorithm based on the measure of edge betweenness. Edges with high edge betweenness are considered to be potential links between communities and are removed one after another to reveal the underlying community structure of a network. Newman (2004b) also propose a faster agglomerative algorithm for optimising modularity, in which

edges whose presence will contribute to the greatest increase or smallest decrease in modularity are added to the network one after another. Comparisons of different algorithms for optimising modularity can be found in (Danon et al., 2005).

Considering both the efficiency and performance of the above algorithms, we choose to use the fast greedy algorithm that has been extended to handle weighted networks (Newman, 2004a) to perform clustering on the networks induced from a folksonomy. It should be noted that the purpose of our experiments is to understand the differences between different networks induced from a folksonomy in revealing the multiple meanings of ambiguous tags, instead of investigating which clustering algorithm is best at performing the task. Hence, while it is true that different algorithms would probably produce different clustering results, we believe the algorithm we have chosen will give us qualitative insights into the differences between the networks that can be generalised to other situations.

4.6.2.2 Clustering of Folksonomy Networks

By tag contextualisation we mean the process of finding out the different contexts in which a tag is used. The result of such process will be one or more sets of tags which when presented with the tag in question point to different concepts it represents. Given the networks described in Section 4.6.1, a network clustering algorithm is expected to return a set of clusters, with each of them hopefully corresponds to one context in which a tag is used. The process involves the following three steps.

1. Firstly, we construct either a network of documents or tags based on one of the methods mentioned in the previous section. This is represented as an adjacency matrix.

2. Secondly, we apply the clustering algorithm to the network and obtain a set of clusters of nodes.

3. Thirdly, we extract labels for each of the clusters. For document networks, we extract the N most popular tags among the documents in a cluster. For tag networks, we extract the top N tags which are most frequently used with the tag in question. These tags constitute the different contexts the clusters correspond to.

As an illustrating example, consider the tag `wine`. We observe that the tag has been used by users in Delicious in two different contexts: (1) as a kind of alcoholic drinks and (2) as the name of a software application. The clustering process performed on the tag context similarity network of `wine` returns two clusters, with one corresponding to the first context and another corresponding to the second. The top five tags extracted as labels for the two clusters are:

1. {food, shopping, shop, drink, vino}
2. {linux, ubuntu, howto, emulation, windows}

Hence, although this method of extracting sets of tags as labels for the clusters does not produce exactly the different meanings of a tag, the most frequently used tags in a cluster actually constitute a coherent context from which the exact meaning of the tag can be easily deduced. This form of representation also facilitates further utilisation of the information in other applications in which comparisons between sets of tags are required.

4.6.3 Experiments

Our experiments are conducted using the data sets collected from Delicious described in Chapter 3. The aim of this experiment is to find out which types of network induced from a folksonomy best capture the associations between the users/documents/tags such that the social meanings of the tags under investigation can be identified. Evaluation of tag contextualisation is challenging due to a lack of a 'gold standard' or a ground truth. Without a thorough understanding of what every user uses a tag to mean (which is by itself impractical and quite impossible), we would not know how many different meanings a tag has within a folksonomy.

Similar studies in word sense discrimination usually resolve to a small set of manually-examined samples, or to the use of pseudowords—artificially ambiguous words created by combining two different words together (Schütze, 1998). The use of pseudowords is not suitable in our case as the user groups of two different tags may be very different such that results may not be useful in general. In addition, the use of an established dictionary such as WordNet (Miller, 1995) as a ground truth may also not be as helpful as one would expect. This is because it is very possible that not all meanings of a tag defined in the dictionary are used by users in Delicious, and there may be new meanings of the tag which do not necessarily

Tag	Context	Label
architecture	physical structures	design, home, art, travel, urban
	programming	design, software, reference, development, webdev
bridge	networking	networking, network, wifi, wireless, linux
	card game	games, cardgame, poker, resources
	architecture	architecture, structure, travel, photos, blog
language	human	reference, education, learning, english, dictionary
	computer	programming, research, reference, software, miscrosoft
opera	music	music, classical, tickets, theatre, woman
	browser	browser, web, software, tools, javascript
sf	science fiction	scifi, fiction, science, literature, sci-fi
	city	sanfrancisco, san, francisco, bayarea, california
soap	cleaning agent	soapmaking, diy, recipes, making, organic
	web services	webservices, programming, xml, web, soa
sun	computer company	solaris, java, linux, programming, unix
	astronomy	science, astronomy, space, photography, solar
tube	video sharing	video, youtube, you, videos, web2.0
	electronics	diy, amplifier, audio, electronics, amp
	underground	london, travel, transport, map, uk
wine	beverage	food, drink, cooking, alcohol, shopping
	software	linux, ubuntu, software, windows, tools
xp	operating system	windows, software, computer, tools, microsoft
	programming	software, development, extremeprogramming, process, agile

TABLE 4.1: Results of the manual classification process. The names of the context are added by us for easier comprehension of the list. The top five tags are shown for each context.

appear in the dictionary. In view of these difficulties, we rely on a small set of manually classified data and perform both quantitative and qualitative analyses to study the characteristics of the different types of network under consideration.

4.6.3.1 Data Preparation

As we have mentioned in Chapter 3, out of the 135 tags chosen as seed tags for data collection from Delicious, 35 tags are selected because they are observed to be used in Delicious in different contexts to refer to different things. We complement this set of tags with 15 randomly selected tags from the remaining 100 seed tags, resulting in a total of 50 tags for this experiment. We have to limit the number of tags examined in this experiment because we have to rely on some human users to manually classify the bookmarks of these tags to establish a basis for comparing the networks.

We ask 10 users who have basic understanding of collaborative tagging systems to classify documents randomly chosen from the data sets by examining the intended meaning of a specific tag. For example, with respect to the tag `sf`, a participant would put documents about San Francisco in one group, and those about science fiction in another. Each user examines the data of two tags, each containing 50 randomly selected documents. Hence every dataset is examined by two participants. We obtain the final outcomes by combining the classifications given by two

participants, i.e. obtaining their consensus. From these 50 tags, we select 10 tags of which the classifications of the two participants most agreed on, i.e. having two or more common contexts and two or less different contexts. These tags represent a good range of topics from different domains. We use sets of tags extracted from different groups of classified documents as their labels. Table 4.1 gives the result of this manual classification process.

While the manual classification process does not necessarily return all the contexts in which the selected tags are used, they do provide a reasonably good common ground for the comparison of the different networks described in the previous section. In the following section, we describe the performance measures used in our quantitative analysis.

4.6.3.2 Performance Measures

To evaluate the results of tag contextualisation, we need some performance measures. Firstly, we introduce several mathematical notations that would facilitate discussion of the experimental results. We denote the set of contexts discovered automatically by the clustering algorithm by $\mathbf{S}_t^A = \{s_{t,i}^A\}$, and the set of manually discovered contexts by $\mathbf{S}_t^M = \{s_{t,i}^M\}$. In addition, we define a match function which, given the set $\mathbf{S}_t^A$ of automatically discovered contexts and a particular manually discovered context $s_{t,i}^M$, returns the number of automatically discovered contexts which match the manually discovered one:

$$\text{match}(\mathbf{S}_t^A, s_{t,i}^M) \tag{4.19}$$

It should be noted that the function does not compare directly two sets of contexts, because it is possible that two contexts returned by the clustering algorithm correspond to the same manually discovered context, a situation we refer to as redundancy which will be further explained below.

We introduce two performance measures here which will be used to study the differences between the aforementioned networks. Note that we do not measure the precision of the contextualisation process. This is because we do not really have a clear idea of what is an incorrect outcome. Given the limited data in the manual classification process, the clustering process is very likely to discover contexts that have not been identified in the former. Hence, we believe it would be more useful to study the following two measures, namely recall and redundancy, as well as to qualitatively look into the results to see if unexpected contexts are

meaningful ones.

$$\text{Recall} = \frac{|\{s_{t,i}^M | \text{match}(\mathbf{S}_t^A, s_{t,i}^M) > 0\}|}{|\mathbf{S}_t^M|} \tag{4.20}$$

$$\text{Redundancy} = \frac{\sum_{s_{t,i}^M \in \mathbf{S}_t^M} F(\mathbf{S}_t^A, s_{t,i}^M)}{|S_t^A| - 1} \tag{4.21}$$

where

$$F(\mathbf{S}_t^A, s_{t,i}^M) = \begin{cases} \text{match}(\mathbf{S}_t^A, s_{t,i}^M) - 1 & \text{if match}(\mathbf{S}_t^A, s_{t,i}^M) > 1 \\ 0 & \text{otherwise} \end{cases} \tag{4.22}$$

As the equation suggests, ***recall*** measures the fraction of contexts discovered by the automatic process with respect to the contexts which are manually discovered. High recall means that the automatic process is able to discover more contexts in which a tag is used. Therefore, the result can be considered better if the level of recall is higher. ***Redundancy***, on the other hand, measures how many clusters returned by the clustering algorithms actually correspond to the same context. Higher redundancy means that extra effort is needed to combine similar contexts. A good result should achieve high recall (returning all the contexts discovered by the manual process), and low redundancy (all contexts returned are unique).

4.6.3.3 Quantitative Analysis

We apply the chosen clustering algorithm to each of the different types of network for the 10 selected tags. We manually calculate the recall and redundancy measures by examining the tags extracted from the clusters. In most cases, the tags alone clearly reveal what the contexts are, with large overlap with the tags extracted in the manual process. There are also cases in which the tags constitute contexts which are not discovered in the manual process. The results are summarised in Table 4.2 and Figure 4.4.

Tag	TD				UD				TC				TC'				CS				CS'			
	N	Rl	Ry	E	N	Rl	Ry	E	N	Rl	Ry	E	N	Rl	Ry	E	N	Rl	Ry	E	N	Rl	Ry	E
architecture	5	1.0	0.6	0	3	1.0	0.3	0	6	1.0	0.3	1	8	1.0	0.3	1	2	1.0	0.0	0	2	1.0	0.0	0
bridge	3	0.8	0.0	0	14	1.0	0.6	0	6	0.5	0.3	0	7	1.0	0.3	0	2	0.5	0.0	0	4	0.8	0.0	0
language	3	1.0	0.3	0	6	1.0	0.7	0	7	1.0	0.7	0	8	1.0	0.8	0	3	1.0	0.3	0	2	1.0	0.0	0
opera	3	1.0	0.3	0	9	1.0	0.8	0	5	1.0	0.6	0	7	1.0	0.7	0	3	0.5	0.3	0	3	1.0	0.3	0
sf	2	1.0	0.0	0	8	1.0	0.6	1	6	1.0	0.7	0	7	1.0	0.4	0	2	1.0	0.0	0	3	1.0	0.3	0
soap	5	0.5	0.6	0	11	1.0	0.7	1	9	1.0	0.8	0	8	1.0	0.6	0	3	1.0	0.3	0	4	1.0	0.0	1
sun	3	1.0	0.3	0	10	1.0	0.8	1	9	1.0	0.7	0	8	1.0	0.6	0	4	1.0	0.5	1	3	1.0	0.3	0
tube	3	1.0	0.0	0	8	1.0	0.6	0	5	0.7	0.0	0	11	1.0	0.7	0	4	1.0	0.3	0	3	0.0	0.0	0
wine	3	1.0	0.3	0	12	1.0	0.8	0	3	1.0	0.3	0	5	1.0	0.4	0	2	1.0	0.0	0	2	1.0	0.0	0
xp	3	1.0	0.3	0	7	1.0	0.7	0	6	1.0	0.7	0	7	1.0	0.7	0	3	1.0	0.3	0	2	1.0	0.0	0

TABLE 4.2: Results of the tag contextualisation process. The network types are TD (tag-based document network), UD (user-based document network), TC (tag co-occurrence network), TC' (tag co-occurrence network with user information), CS (tag context similarity network), and CS' (tag context similarity network with user information). N stands for number of clusters, Rl for recall, Ry for redundancy, and E is the number of extra contexts discovered.

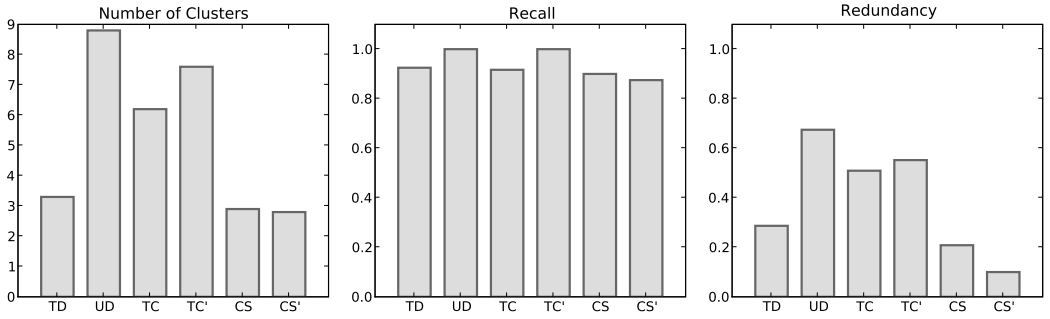


FIGURE 4.4: Average number of clusters, recall and redundancy of the tag contextualisation process.

Figure 4.4 shows that the user-based document networks (UD) and tag co-occurrence networks (TC and TC') produce the largest number of clusters. By comparing the tag-based document networks (TD) and UD, we find out that edge weights in TD are usually higher than in UD. This is because documents sharing a similar set of tags may not be tagged by a similar set of user. In particular, there can be no edges between two documents if there are no users who have assigned the tag t to both of them, even though they are about a similar topic. Hence, the clustering algorithm breaks down UD into more and smaller clusters. For TC and TC', the relatively large number of clusters can be explained by the fact that many documents share only a few popular tags, and the other tags thus are less connected with each other, forming small groups of tags in the networks.

The tag context similarity networks (CS and CS') return the fewest number of clusters. This is because they do not only incorporate co-occurrence information but also involve the comparison of the contexts of each tag in calculating their similarity, which is also known as second-order co-occurrence (Schütze, 1998). This is actually similar to the idea of latent semantic analysis (Deerwester et al., 1990). In other words, tags are not connected only because they have been used together directly, but because they have been used with other similar tags (having similar contexts). This increases the number of edges and edges weights in a tag network, thus vertices are more connected with each other, resulting in a smaller number of clusters.

Recall, on the other hand, is generally high in all of the cases. In particular, both UD and TC' achieve 100% recall. In fact, the manually discovered meanings of the tags can be identified for most of the time except in the cases of a few tags such as `bridge` and `tube` which have more different meanings than the other tags.

A closer look at the clusters in CS and CS', which achieves relatively lower recall, reveals that tags related to the missing meanings are included in a cluster which corresponds to a different meaning. For example, in the case of CS for the tag `bridge`, tags related to architecture and those related to networking are mistakenly grouped under the same cluster. This means that the context similarities between some less related tags are too strong such that the clustering algorithm is unable to split them into two groups.

The bar chart of redundancy levels has a similar shape as that of number of clusters. This is because when the clustering algorithm returns more clusters it is more likely that two or more clusters correspond to the same context, especially when the number of contexts in which a tag is used in Delicious is limited. However, we also note that high redundancy levels in some cases are also due to the fact that the contexts discovered in the manual classification process are too general, such that some more specific contexts discovered in the clustering process are mapped to the same contexts. We will discuss more about this in the next section.

Redundancy is important. This is because when redundancy is too high the results can not be directly used in other applications. Some post-processing steps will be needed to combine clusters which correspond to the same context. Given that we label the clusters with sets of tags extracted from the clusters, one way to combine the clusters is to compare the sets of tags and perform a merge if there is significant overlap. The second option is to filter away clusters of small size. For example, if we remove clusters of size less than 5% of the total number of nodes in the user-based document networks (UD), it achieves a redundancy level of 0.3, similar to that of TD, while maintaining a recall level of 1.0.

4.6.3.4 Qualitative Analysis

Firstly, we look at the extra contexts discovered by the clustering algorithm. The use of UD returns the largest number of 'new meanings' of the tags we examine. For example, it reveals that the tag `sf` is also used by Delicious users to refer to 'Sourceforge', an open source software repository on the Web. It also reveals that the tag `soap` is also used to refer to 'TV dramas'. These meanings are not identified by tag-based networks such as TD and TC. A closer look at the documents and tags in the corresponding clusters reveals that only a relatively small number of users are using the tags for those meanings (about 5% of users for 'Sourceforge' and less than 2% for 'TV dramas').

If we perform clustering based on tags, tags which are used in those contexts are likely to be mixed up with other tags if they co-occur in some other documents. On the other hand, by connecting documents based only on the users (as in the case of UD), it is more likely that documents which are about the same topic would be grouped together, causing also tags used in the same context to be grouped together as well.

In addition, the existence of subtopics among the clusters is another aspect which is not reflected in the quantitative performance measures. The meanings discovered in the manual classification process (Table 4.1) are actually rather general. For example, while `sun` is found to be used to refer to the computer company, there are some clusters which point to particularly the Java programming language developed by the company, and some others which point to the company's Solaris operating system. In this respect, clustering of UD returns more subtopics than the other networks. For example, for the tag `sf`, there are clusters about food and restaurants in San Francisco, while others are about hiking and outdoor activities in the city. For the tag `language`, there are clusters which correspond to different languages such Chinese, English and Japanese. The context similarity networks CS and CS', which return the least number of clusters, return the least number of subtopics. This is probably because the context similarity tends to group tags into as general groups as possible.

Finally, we also notice that clustering on UD also returns some language-specific clusters. There are clusters with only Chinese documents described mainly by Chinese tags, and some others with only Japanese documents described mainly by Japanese tags. This suggests that user-based networks are also able to identify specific user communities of different languages in a collaborative tagging system.

4.6.3.5 Comparison with Ontologies

We compare the contextualisation results with the meanings returned by WordNet (Miller, 1995). WordNet is an English lexicon which groups words into sets of synonyms called synsets. It also distinguishes between different senses of a polysemy word by associating the word with different synsets. By querying WordNet, it is possible to find out the different meanings of an ambiguous word.

We submit each of the 10 tags as queries to the online interface of WordNet.[5] Each query returns a set of synsets in which the tag appears. For example, submitting

[5] `http://wordnetweb.princeton.edu/perl/webwn`

a query to WordNet using the tag `opera` would obtain the following: (1) a drama set to music, consists of singing with orchestral accompaniment and an orchestral overture and interludes; (2) a commercial browser; and (3) opera house, a building where musical dramas are performed.

We manually compare these synsets with the results of clustering described in the previous section. It should be noted that it is not trivial to perform the comparison as there may not be a one-to-one mapping between the contexts discovered in the clustering process and the synsets returned by WordNet. For example, WordNet returns four synsets for the tag `architecture`, three of which are related to building physical structures and refer to the structures, the discipline and the profession respectively. When performing the comparison, we consider the above three synsets to be all matched by one of the contexts discovered in the clustering process.

We are aware of several important discoveries in this qualitative study of the results of the tag contextualisation process. Firstly, the number of synsets returned by WordNet for a tag is usually larger than that of contexts discovered in the clustering process. WordNet also returns more fine-grained results. For example, *soap* can be used to refer to money offered as a bribe, and is used as the street name of a drug. While this suggests that the the contexts discovered in the clustering process may not be comprehensive enough, it is also possible that these additional meanings of the tag are never or rarely used in Delicious. If the aim of tag contextualisation is to enhance organisation and retrieval of documents in tagging systems, additional meanings of a tag would not be very useful.

A more important finding is that quite a number of contexts discovered in the clustering process cannot be found in WordNet. These include 'programming' (architecture), 'networking' (bridge), 'web services' and 'TV dramas' (soap), 'computer company' (sun), 'video sharing' (tube) and 'software' (wine). In addition, WordNet does not offer any information about abbreviations such as *sf* and *tube*. One may suggest that these meanings can be found on Wikipedia's disambiguation pages. However, Wikipedia offers mainly textual information and it is difficult to query Wikipedia for structured data (DBpedia (Auer et al., 2008) and YAGO (Suchanek et al., 2007), which are attempts to construct ontologies by extracting information from Wikipedia, do not contain all the disambiguation information). In addition, the disambiguation pages of Wikipedia usually contain a lot of meanings of a term, most of which are not found to be used in Delicious.

In summary, this analysis suggests that querying external resources may not be a

suitable way of obtaining the different contexts in which ambiguous tags are used. In this respect, unsupervised clustering methods are more suitable especially when we want to find out how the tags are actually used within a collaborative tagging system.

4.7 Discussion

The above experiments and analysis on tag contextualisation suggest that the use of graph-based clustering algorithms to perform tag contextualisation on an individual tag level produces promising results. Our findings can be summarised as follows.

- Tag-based document networks, while being one of the simplest forms of network derived from a folksonomy, do not favour the identification of meanings used by only a small number of users or a specific user group.
- Tag context similarity networks tend to capture the most general concepts represented by the tags being disambiguated. It provides the most clear-cut results among all the network types. However, it also tends to miss some contexts in some cases.
- User-based document networks facilitate the identification of many subtopics with are actually interesting to the users in the folksonomy, and it even helps to identify user communities with respect to a particular topic. This is probably due to the fact that these networks ground the relationship between documents on the social context, i.e. the group of users who are interested in them.
- Automatic clustering of folksonomy networks for tag contextualisation produces satisfactory results. Compared to the use of external resources such as dictionaries and ontologies, it is more likely to identify the different contexts in which the tags are actually used within the system.

In fact, while tag contextualisation has its own merits in revealing the social meanings of the tags in a collaborative tagging environment, the technique has many other applications on the Web. For example, the identified contexts as well as the corresponding relevant tags can be used directly to classify documents in a folksonomy. When a user searches for documents with the tag `sf`, the system can use

the sets of tags which correspond to the different contexts of the tag to partition the result into two or more groups of documents of different topics, thus facilitating the user in locating documents most relevant to his needs. In the following section, we present our experiments based on the idea of classifying Web search results returned by the popular Web search engine Google by using results of tag contextualisation.

4.8 Web Search Result Classification

Due to the huge volume of resources available on the Web, searching information which is relevant and useful has become more and more difficult. Web search engines such as Google and Yahoo! are designed to present resources which satisfy the information needs of Web users. However, information needs become less clear once they are translated into queries composed of individual keywords, which are still the dominant type of input in Web search. This becomes a problem particularly when the keywords are polysemous, i.e. they represent different concepts depending on the contexts in which they are used. We discuss this problem by describing an example and by mentioning some previously proposed solutions.

4.8.1 Query Ambiguity and Web Search Classification

Let us look at an example which illustrates the problem of keyword ambiguity in Web search. Consider the situation in which a user wishes to search for information about contract bridge (a card game) on the Web. When the user submits a query with the keyword *bridge* to the search engine Google, a list of documents is returned. While in the ideal case all documents should be relevant to the card game which is desirable from the perspective of the user, it is usually not the case in practise. As search engines commonly adopt the keyword-based retrieval approach, documents containing the query term *bridge* are all likely to be returned. As *bridge* carries several meanings, it is not surprising that the documents returned are about very different topics.

Table 4.3 shows the top ten documents returned by Google.[6] We can see that while several items in the list are about the card game, it also contains documents that

[6]This list would change as Google re-calculates the PageRank of the documents it has indexed. However, the problem of ambiguity is likely to persist due to the fact that semantics of a keyword is not taken into account in the query process.

Rank	Title	URL
1	Bridge - Wikipedia, the free encyclopedia	http://en.wikipedia.org/wiki/Bridge
2	Contract bridge - Wikipedia, the free encyclopedia	http://en.wikipedia.org/wiki/Contract_bridge
3	American Contract Bridge League - Home Page	http://www.acbl.org/
4	Bridge: rules and variation of the card game	http://www.pagat.com/boston/bridge.html
5	Bridge - Mainstreaming Gender Equality	http://www.bridge.ids.ac.uk/
6	Bridge Base Online	http://www.bridgebase.com/
7	media manager, file browser — Adobe Bridge CS4	http://www.adobe.com/products/creativesuite/bridge/
8	Bridge Ocean Education Teacher Resource Center	http://www.vims.edu/bridge
9	Bridge Records, Inc.	http://www.bridgerecords.com/
10	Play bridge card game online	http://www.bridgeclublive.com/

TABLE 4.3: The top ten documents returned by the Google search engine when *bridge* is used as a query term.

address other meanings of the word *bridge*. For example, the first document is a page from Wikipedia describing bridges as architectural structures. There are also documents (e.g. 7th, 8th and 9th) that contain information about organisations or projects named *Bridge* but are by no means related to any commonly used meanings of the word.

From this example, two major problems can be observed. Firstly, extra effort is required from the user to go through the list and single out documents which are relevant to his/her information needs. In this case, the user needs to check whether each of the returned documents is about the card game called bridge. Secondly, the presence of irrelevant documents reduces the number of relevant documents which can be presented to the user at one time. This is particularly important because it is found that users tend to inspect only the first set of documents returned (Silverstein et al., 1999; iProspect, 2006).

It should be noted that while a user can make a query more specific by adding other keywords to the query string to narrow down the search result, single-term queries are found to be very common, representing 20–35% of all queries according to several Web search studies (Jansen et al., 2000). Even though some search engines such as Google provide suggestions to users on how they can refine the search results by presenting potentially related keywords, it is still more desirable from the perspective of the users that the search results are first classified into different categories.

Word ambiguity is studied extensively under the field of word sense disambiguation (Ide and Veronis, 1998), which focuses on developing methods for identifying the sense of an occurrence of an ambiguous word. Word sense disambiguation can

be divided into two different sub-tasks, namely sense discrimination and sense labelling (Schütze, 1998). Sense discrimination divides the occurrences of a word into a number of classes by determining for any two occurrences whether they belong to the same sense. Sense labelling, on the other hand, assigns a sense to each class and to each occurrence of the ambiguous word. A lot of different methods have been proposed for word sense disambiguation, such as using machine readable dictionaries or thesauri (Krovetz and Croft, 1989; Lesk, 1986), or by clustering of keywords in documents based on their co-occurrences (Schütze, 1998).

Document clustering can be considered as a solution to the problem of word ambiguity in Web search from a different perspective. Instead of figuring out the different senses of the query terms, one can perform clustering on the documents returned by a Web search engine, such that each resultant cluster would contain documents which address the same sub-topic of the query (Cutting et al., 1992; Hannappel et al., 1999; Zamir and Etzioni, 1998). Carrot[2] (Stefanowski and Weiss, 2003) is a search result clustering engine powered by several different clustering algorithms.[7] Vivismo[8] (Koshman et al., 2006), Grokker[9] and iBoogie[10], for example, are commercial systems which provide similar functionalities. All these search engines collect, with respect to a query, Web documents returned by other Web search engines, perform clustering on these documents and generate labels for the resultant classes. There are also proposals of employing supervised learning methods to classify documents (Chekuri et al., 2007; Zeng et al., 2004). However, supervised learning methods are less suitable in the context of Web search as it is rare that proper training datasets are available.

4.8.2 Enhancing Web Search using Folksonomies

As we have shown earlier in this chapter, meanings of tags can be discovered by performing clustering on networks induced from folksonomies. Collaborative tagging systems thus represent a valuable source of information for understand the different contexts in which a particular term is used. We therefore believe that the information can be utilised to provide a possible solution to the problem of keyword ambiguity in Web search.

[7] Carrot[2]: `http://project.carrot2.org/`

[8] The public version of Vivismo's Web search engine, Clusty, can be found at `http://clusty.com`.

[9] Grokker: `http://www.grokker.com/`

[10] iBoogie: `http://www.iboogie.com/`

Our proposed method involves two phases. We use our method of tag contextualisation to identify the different contexts in which a particular tag is used. Here, we mainly focus on user-based document networks (UD), because these networks are found to reveal more social meanings of a tag. We aim at building document classifiers based on a clustering process performed on the folksonomy. At the end of the clustering process, we should have two pieces of information: (1) one or more clusters of documents, and (2) a set of class labels in the form of sets of tags.

Secondly, we apply these classifiers to classify documents returned by a Web search engine when the ambiguous tag is used as a query. If we assume that documents returned by a search engine can be represented as a term vector in which elements indicate the weights (importance) of the corresponding keywords, a simple and straightforward approach to classify the documents would be to compare the term vectors which the class labels of the clusters. A document can be put into the class whose label it is most similar to. A threshold can also be specified so that documents which are not sufficiently similar to any of the class labels will not be assigned a class, so as to reduce the chance of false positive cases. This approach is examined in a preliminary paper (Au Yeung et al., 2008e).

However, our initial experiments actually suggest that there exist some problems with this straightforward approach. In particular, the keywords characterising the documents can be more diverse than the tags extracted from Delicious, since Delicious only contains a rather small subset of the documents available on the Web. Hence, a document may not be put into the right group even if it should be, as the similarity values between the term vector of the document and the class labels can be quite low.

To remedy this problem, we instead go for a k-nearest-neighbour classification approach. Recall that one or more clusters of documents are returned by the clustering process. Documents within the same cluster should all be relevant to a particular context in which the tag in question is used. In this sense, document tagged by users in a folksonomy can be regarded as training samples of a k-nearest-neighbour classifier. The term vectors of new documents can be compared with each of these training samples, and can be put into a particular class based on majority vote. In this way, the new document will be classified based on a larger sample of documents instead of only on the popular or important tags extracted from each context.

4.8.2.1 Building Classifiers from Folksonomies

Based on the above considerations, we now present a formal description of our proposed method for Web search result classification. First we describe the process of initialising a classifier of Web search results with respect to an ambiguous keyword t. Recall that an adjacency matrix $\mathbf{B} = \{b_{ij}\}$ representing a user-based document network for the tag t can be constructed by using Equation 4.4, reproduced below:

$$b_{ij} = |\{u|(u,t,d_i) \in R \wedge (u,t,d_j) \in R\}|$$

We again apply the clustering algorithm that optimises modularity to divide the network into different groups of vertices. Recall that D_t is the set of documents which have been assigned the tag t. The result of the clustering process is then a set of clusters of documents:

$$X_t = \{X_{t,1}, X_{t,2}, ..., X_{t,m}\} \tag{4.23}$$

where

$$X_{t,1} \cup X_{t,2} \cup \cdots \cup X_{t,m} = D_t \tag{4.24}$$

Finally, for each cluster $X_{t,i}$, we extract a set $T_{t,i}$ of tags as its class label.

As we have shown in Section 4.6.3.3, this kind of network would produce a result with high redundancy. In other words, while each of these clusters should correspond to a single context in which the tag t is used, it is possible that two or more of these sets refer to the same context. To eliminate such redundancy we combine two clusters if there is significant overlap between the two class labels with the help of the following function:

$$overlap(T_{t,i}, T_{t,j}) = \frac{|T_{t,i} \cap T_{t,j}|}{|T_{t,i} \cup T_{t,j}|} \tag{4.25}$$

With the help of a threshold value we call α, we merge two sets of documents $X_{t,i}$ and $X_{t,j}$ when $overlap(T_{t,i}, T_{t,j}) \geq \alpha$. A new class label is generated for the new cluster. Hence, the final result of this process is a set of classes of documents:

$$\mathbf{C}_t = \{C_{t,1}, C_{t,2}, ..., C_{t,n}\} \tag{4.26}$$

with class labels $\{T_{t,1}, T_{t,2}, ..., T_{t,n}\}$, where $n \leq m$. The classes together represent a k-nearest-neighbour classifier of documents returned by the search engine when

Algorithm 1 Building K-Nearest-Neighbour Classifier from Folksonomy

Input: Adjacency matrix $\mathbf{M}$ of the network of documents
Output: A set $\mathbf{C}$ of classes with a set of labels $\mathbf{T}$

{Document clustering}
$\mathbf{C} \Leftarrow NetworkClustering(\mathbf{M})$
$\mathbf{T} \Leftarrow \{\}$
{Extract frequent tags}
for $C_i \in \mathbf{C}$ **do**
  $T_i \Leftarrow ExtractTags(C_i)$
  $\mathbf{T} \Leftarrow \mathbf{T} \cup \{T_i\}$
end for
{Merge similar clusters}
$merged \Leftarrow 1$
while $merged = 1$ **do**
  $merged \Leftarrow 0$
  for $T_i, T_j \in \mathbf{T}$ and $i \neq j$ **do**
    if $overlap(T_i, T_j) \geq \alpha$ **then**
      $C_{new} \Leftarrow C_i \cup C_j$
      $\mathbf{C} \Leftarrow \mathbf{C} - \{C_i, C_j\}$
      $\mathbf{C} \Leftarrow \mathbf{C} \cup \{C_{new}\}$
      $T_{new} \Leftarrow ExtractTags(C_{new})$
      $\mathbf{T} \Leftarrow \mathbf{T} - \{T_i, T_j\}$
      $\mathbf{T} \Leftarrow \mathbf{T} \cup \{T_{new}\}$
      $merged \Leftarrow 1$
    end if
  end for
end while
return **C**,**T**

Class	Label
1	bridge, programming, development, library, code, ruby, tools, software, adobe, dev
2	bridge, games, cards, game, imported, howto, conventions, card, bidding, online
3	bridge, networking, linux, network, howto, software, sysadmin, firewall, virtualization, security
4	bridge, bridges, structures, engineering, science, physics, school, education, building, reference

TABLE 4.4: Result of clustering on documents tagged with `bridge` in Delicious.

t is used as a query term. The whole process is summarised in Algorithm 1.

As an example, Table 4.4 shows the result of this process when the algorithm is applied to documents tagged with `bridge` in Delicious, with $\alpha = 0.3$. It can be observed that the frequent tags extracted as class labels for the clusters clearly indicate the different contexts in which the tag is used by the users in Delicious.

4.8.2.2 Web Search Result Classification

Given the set of classes obtained in the previous step, we can then apply them to document classification. We assume that a set S_t of documents will be returned by a search engine, when it is queried with a keyword t. Our target is to classify the documents into the different classes obtained from the folksonomy clustering process described in the previous section.

We treat the results obtained in the previous step as k-nearest-neighbour classifiers. When a new document is observed, it is compared with all known documents and their degrees of similarity are calculated. The known documents will be ordered in descending order of their degrees of similarity with the new document. The new document will then be assigned to the class to which the majority of the top k known documents in the list belong to. k is usually chosen to be an odd integer so as to avoid ties.

We assume that each document $s_{t,j} \in S_t$ is characterised by a set $K_{t,j}$ of keywords. This set $K_{t,j}$ can be constructed from keywords extracted from the document text using common information retrieval techniques such as the TF-IDF weighting scheme (Manning et al., 2008). On the other hand, if the document has been posted to a collaborative tagging system, the set of tags assigned to the document in the system can be treated as the set of keywords for this document. Of course, a combination of the two will provide more information about the content of the document. By using this set of keywords, we can then calculate the similarity between such document and a document $d_{t,i}$ in D_t, the set of documents that have been assigned the tag t in a folksonomy and have been clustered into one of the classes in $\mathbf{C}_t$.

Let $J_{t,i}$ be the set of tags assigned to $d_{t,i}$. The similarity measure we use here to compare two sets of keywords of two documents is the Dice coefficient, which is given by the following equation:

$$Sim(K_{t,j}, J_{t,i}) = \frac{2 \times |K_{t,j} \cap J_{t,i}|}{|K_{t,j}| + |J_{t,i}|} \tag{4.27}$$

Based on this similarity measure, a classification process can be summarised as follows. For each $s_{t,j} \in S_t$, we obtain the k most similar documents from the set D_t. The class of $s_{t,j}$ is decided by the majority votes of the classes of these k nearest neighbours. Documents belonging to the same class can then be grouped together before the search result is presented to the user.

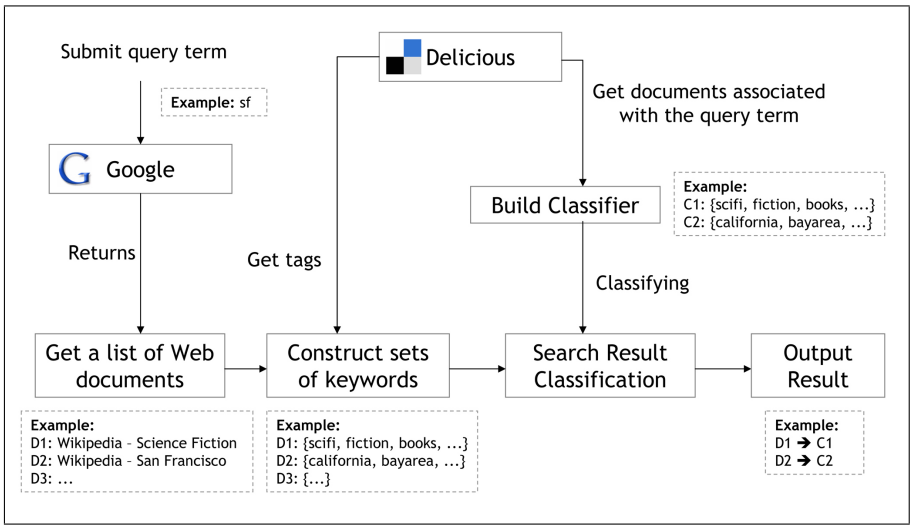


FIGURE 4.5: Flow chart of Web search classification using folksonomies.

It should be noted that while the classes correspond to different contexts in which t is used, the classes cannot be considered as exhaustive of all possible contexts. It is possible that a meaning of t is never referred to by the users in the system, and is not identified by the clustering algorithm. It is therefore possible that a document in the search result cannot be classified into any of the classes in $\mathbf{C}_t$. Hence we introduce a threshold β here, where $0 \leq \beta \leq 1$. For a particular document $s_{t,j}$, if half of the k nearest neighbours have a similarity value less than β, the document will be assigned the class $C_{t,0}$, which represents unclassified documents. Furthermore, we represent this classification process as a function which maps a document to a class with respect to a tag:

$$F_A : S_t \times T \rightarrow \mathbf{C}_t \tag{4.28}$$

The subscript A means automatic classification, in contrast to the manual classification process which will be described in the next section in which our experiments are described.

4.8.3 Experiments

To experiment with this idea of using results of tag contextualisation for Web search result classification, we apply the method to results returned by the Google search engine for the ten tags we have examined in Section 4.6.3: `architecture`, `bridge`, `language`, `opera`, `sf`, `soap`, `sun`, `tube`, `wine`, `xp`. Figure 4.5 and Figure 4.6 show an overview of the process of Web search classification.

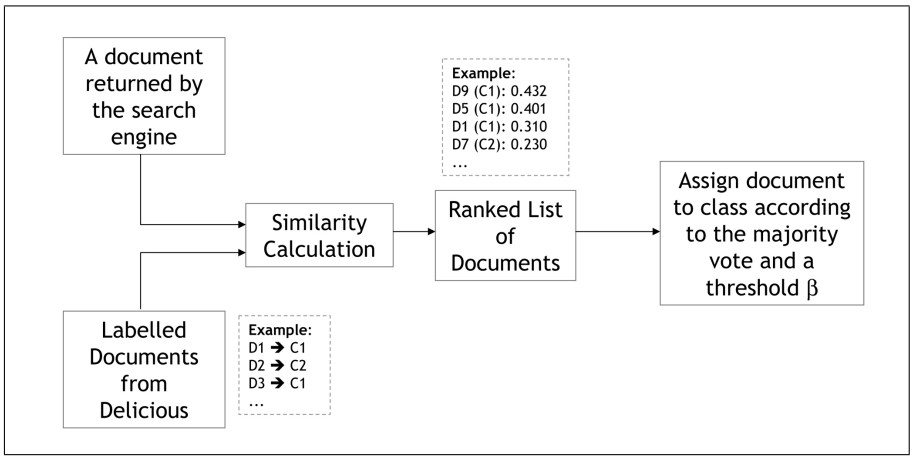


FIGURE 4.6: Classification of documents returned by a Web search engine.

4.8.3.1 Experimental Setup

To build the k-nearest-neighbour classifiers, we run our algorithm for building classifiers on the ten user-based document networks induced from the data sets of the ten tags. For the testing data set, we submit queries using each of the ten terms to the search engine Google and collect the first 50 documents returned. We then extract all the tags assigned to these documents in Delicious. For some documents that are not assigned any tags in Delicious, we extract keywords from the content of the documents by removing common stop-words. Among the 500 documents collected from Google, 469 or 93% of the documents are found to be assigned tags in Delicious. We denote the set of documents retrieved for the term t by S_t.

We first apply the clustering process to obtain a set of classes for each of the tags. We choose $\alpha = 0.3$ such that it requires an overlap of about half of the tags in two class labels for two clusters to be merged. In practice, if two clusters contain documents addressing the same meaning of a tag their class labels are very similar to each other. We also do not see any errors in the step of combining clusters for different tags. The above choice of the value of α is therefore a reasonable one.

We extract the 10 most frequently used tags from each resultant class of documents. These 10 tags are treated as class labels. The result is shown in Table 4.5. The names of the contexts are added by us for better understanding of the classes. It can be observed that the proposed algorithm performs well in revealing the different contexts in which the tags are used. The tags extracted are also closely related to the contexts they represent.

Next, we apply our k-nearest-neighbour classification method to search results

Tag	Context	Tags Extracted
architecture	Software design	architecture, design, programming, toread, development, software, article, work, reference, web
	Buildings design	architecture, design, art, inspiration, cool, home, blog, house, culture, arquitectura
bridge	Design pattern	bridge, programming, development, library, code, ruby, tools, software, adobe, dev
	Card game	bridge, games, cards, game, imported, howto, conventions, card, bidding, online
	Computer networking	bridge, networking, linux, network, howto, software, sysadmin, firewall, virtualization, security
	Architecture	bridge, bridges, structures, engineering, science, physics, school, education, building, reference
language	Human language	reference, english, learning, dictionary, education, languages, writing, tools, translation
	Computer language	programming, tutorial, development, functional, software, code, statistics, erlang, java
opera	Web browser	opera, browser, web, software, javascript, tools, tips, internet , browsers, firefox
	Musical performance	opera, music, musique, classical, culture, art, travel ,nyc, musica, classic
sf	San Francisco	sf, sanfrancisco, bayarea, san, francisco, california, travel, events, art, san_francisco
	Science fiction	sf, scifi, fiction, books, sci-fi, literature, writing, sciencefiction, science, fantasy
soap	Cleaning agent	soap, soapmaking, diy, recipes, crafts, shopping, making, beauty, howto, craft
	Web services	soap, webservices, webservice, programming, web, xml, soa, development, wsdl, java
sun	Company	sun, java, programming, opensource, development, solaris, j2ee, web, javafx, software
	Celestial object	sun, astronomy, technology, science, space, moon, solar, education, news, sunrise
tube	YouTube videos	tube, youtube, video, funny, videos, fun, cool, music, feel.good, flash
	Vacuum tubes	tube, audio, electronics, diy, amplifier, amp, tubes, music, elect, guitar
	London underground	tube, london, underground, travel, transport, maps, map, uk, subway, reference
wine	Software application	wine, linux, ubuntu, howto, windows, software, tutorial, emulation, reference, games
	Beverage	wine, food, shopping, drink, reference, vino, cooking, alcohol, blog, news
xp	Windows XP	xp, windows, software, tools, pc, computer, tech, winxp, microsoft, windowsxp
	Extreme programming	xp, software, programming, process, methodology, development, agile, tech, extremeprogramming, extreme_programming

TABLE 4.5: Classes returned by the clustering process for each of the ten tags.

we obtained from Google. In order to evaluate the performance of our proposed method, we have to establish a ground truth against which our result can be compared. Hence, we first manually classify the returned documents into the classes discovered in the clustering process. For example, for the tag `sf`, we have two classes: $C_{\texttt{sf},1}$, which corresponds to 'San Francisco' and $C_{\text{sf},2}$, which corresponds to 'science fiction'. We manually assign each of the documents returned by Google to one of these two classes. If a document cannot be classified to any of the available classes, we assign it the class $C_{t,0}$, which is reserved for unclassified documents. We represent this manual classification as a function which maps a document to a class with respect to a certain tag:

$$F_M : S_t \times T \rightarrow \mathbf{C}_t \tag{4.29}$$

Given the classification functions F_A and F_M, it becomes possible to investigate the performance of our proposed method. We employ three different performance measures here, namely ***precision***, ***recall*** and ***coverage***. Precision measures the extent to which the documents that can be classified are correctly classified. It is calculated by dividing the number of correctly classified documents by the total number of classified documents:

$$P = \frac{|\{d \in S_t | F_M(d,t) = F_A(d,t) \wedge F_A(d,t) \neq C_{t,0}\}|}{|\{d \in S_t | F_A(d,t) \neq C_{t,0}\}|} \tag{4.30}$$

Note that we define $P = 1$ when no documents are classified. Recall measures the fraction of classifiable documents that the method is able to classify:

$$R = \frac{|\{d \in S_t | F_M(d,t) = F_A(d,t) \wedge F_M(d,t) \neq C_{t,0}\}|}{|\{d \in S_t | F_M(d,t) \neq C_{t,0}\}|} \tag{4.31}$$

By classifiable documents we refer to documents which should fall into any one of the contexts discovered in the clustering step. Finally, coverage measures how many documents can be classified given the total number of documents returned:

$$C = \frac{|\{d \in S_t | F_M(d,t) = F_A(d,t) \wedge F_M(d,t) \neq C_{t,0}\}|}{|S_t|} \tag{4.32}$$

We run our experiment using different values of k (number of nearest neighbour) and β (similarity threshold). We find that when $k \geq 5$, the value of k does not have significant effect on the performance of the classification, hence we choose $k = 11$ for the rest of our experiments. Figure 4.7 shows the performance of our

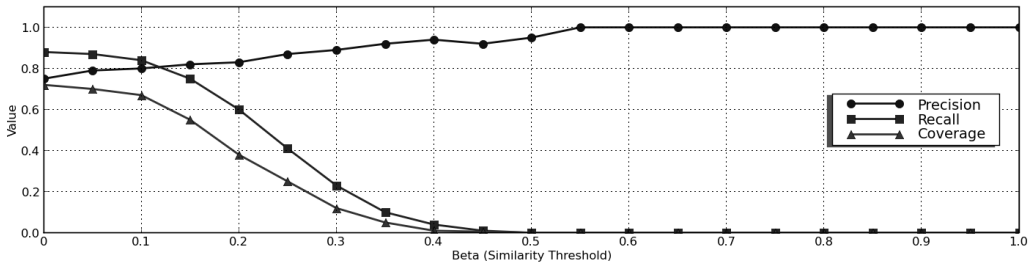


FIGURE 4.7: Precision, recall and coverage against different values of β.

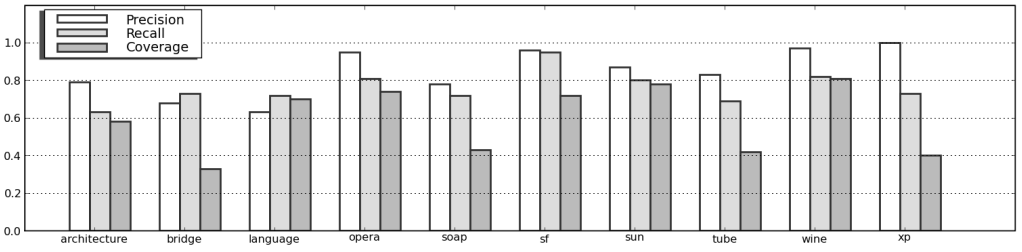


FIGURE 4.8: Precision, recall and coverage for different tags ($k = 11$, $\beta = 0.15$).

method for different values of β. In addition, we also take a closer look at the performance of our proposed algorithm on different tags at $\beta = 0.15$ (Figure 4.8), a value chosen in the range where recall and coverage are high enough for the measure of precision to be meaningful.

4.8.3.2 Results

Figure 4.7 shows that as β becomes larger precision first increases, then decreases, and finally becomes flat at the value of 1.0. The increase of the precision in the first part is reasonable, because misclassified documents are eventually excluded due to their low similarity to the training data as β increases. However, as recall approaches zero, precision is greatly affected when one or two documents are misclassified, thus resulting in the fluctuation. When β becomes so large that no documents are classified, precision becomes one towards the end as we have defined (precision equals one when no documents are classifiable by the classifier). On the other hand, our experiment shows that recall and coverage decrease as β increases. This is because, when β increases, more documents will be considered as unclassified. It should be noted that the calculations of recall and coverage are only different from each other in the denominators, which are constants in both cases. Hence it is not surprising to see that they have similar declining curves.

Overall, the figures show that our proposed method gives satisfactory results. For example, at $\beta = 0.15$, both precision and recall are about 80%, meaning that about 64% of the documents are correctly grouped together under suitable classes. The fact that this result is obtained at a relatively low value of the similarity threshold suggests that the class of an unseen document can be determined when only a few important keywords are present in the metadata or are identified in the content of the document. In other words, the experiment suggests that information required to contextualise an ambiguous tag can usually be found even within a small number of highly related tags.

As a closer look at the experimental results, Figure 4.8 shows the performance measures for different tags at $\beta = 0.15$. Our proposed method gives satisfactory results as judged from the precision of the classifications, ranging from 62% (`language`) to 100% (`xp`). This suggests that the clustering process performed on the folksonomy is able to place documents into meaningful clusters, such that these documents provide an accurate basis for the k-nearest-neighbour classification process. Our investigation into cases with relatively low precision (e.g. `architecture`, `bridge` and `language`) reveals that while misclassified documents contain keywords that provide enough information about their contexts, they are not always the same as the tags assigned to the documents that corresponding to the same contexts in Delicious. For example, documents about bridges as an architectural structure returned by Google contain keywords such as 'river' and 'stream', however these keywords do not appear as tags in Delicious on related documents. This, however, tells us that users in Delicious assign tags they think suitable but do not appear in the content of the documents.

Recall and coverage are relatively lower than precision in all cases. Low recall means that the algorithm is unable to classify many documents that are actually related to one of the contexts discovered in the clustering process. This is probably due to the same reason mentioned above that causes low precision in some cases. Documents in Delicious and those returned by Google are characterised by different keywords of the same context in some cases, resulting a certain amount of documents unclassifiable.

The measure of coverage has the greatest range among the three, with values between 32% (`bridge`) and 81% (`wine`). While low coverage is partly due to low recall in some cases, the result also suggests that the clustering process do not always return all the contexts in which the tag in question is used. For example, the common use of `tube` to refer to a hollow, long and circular structure is not found

in the list of contexts discovered, and because of this some documents returned by Google that describe different kinds of tubes become unclassifiable. However, a further investigation of the documents reveal that the major reason of low recall and coverage is that the unclassified documents are actually not related to any commonly known meanings of the tags in question. For example, search result for `bridge` contains documents about entertainment venues or organisations whose names contain the word `bridge`, but nevertheless have nothing to do with any conventional meanings of the word 'bridge'. Such items account for more than half of the 50 documents we have examined (ignoring these items thus results in over 70% coverage for `bridge`). Judging from this observation, we believe that a low coverage is not as undesirable as it first seems, because our proposed method actually helps to filter out documents that are not semantically related to the query term.

Our proposed method involves first performing clustering documents in a folksonomy based on user-contributed metadata, and then classifying search results returned by Web search engines. While this method resembles traditional machine learning processes, which involve first training a model with classified documents and then applying such model to classify previously unseen documents, there are several major differences.

Firstly, our first step is actually an automatic clustering step which does not require any human input such as labelling a training set of documents. We rely on the user-contributed metadata as well as the community of users to produce meaningful clusters of documents.

Secondly, our method focuses mainly on the tags (metadata) of the documents, only referring to the content of the documents when there are insufficient data found in a tagging system. Tags are user-contributed metadata that provide valuable information about how the users think that the documents should be categorised. They also represent a form of abstraction of the content of the documents. Hence, performing clustering on tag should produce results that are more meaningful than on automatically extracted keywords from a document, which involves estimations of the importance of keywords to the documents. For example, submitting the term `sf` to the Carrot^2 clustering search engine will result in a large number of clusters, in which some are closely related to each other and some may be too trivial to be considered as a separate cluster. Of course, these clustering engines have their merits, such as higher recall and coverage. Therefore how to combine user-contributed metadata and traditional document content analysis to

enhance Web search is a major research question in the near future.

In summary, our proposed method for Web search result classification is able to classify documents with high precision based on the implicit semantics extracted from a collaborative tagging system. Clearly, from the experimental results we believe there are several ways in which the proposed method can be improved. In particular, how we can build a more comprehensive classifier—both in terms of the keywords characterising the documents and of the contexts in which the ambiguous terms are used—is a major issue that requires further investigation.

4.9 Chapter Summary

In this chapter, we have presented a study of the semantics of tags in collaborative tagging systems. The meanings of a tag used by the users, as we have explained, do not always conform to existing and conventional definitions of the word. More likely, the meanings of a tag depend on the community of users who have actually used the tag on some documents. We refer to these meanings of a tag its ***social meanings***.

Preliminary studies show that by analysing the collective user behaviour it is possible to identify clusters of users/documents/tags that correspond to the different contexts in which a tag is used. We also compare different network representations of data in a folksonomy to study which types of network are most useful in revealing the social meanings of the tags. Our result indicates that networks that explicitly take the social context—the users' tagging behaviour and associations actively created by them—into account give a better picture of the semantics of a tag, providing strong support to our hypothesis that we laid out at the beginning of this chapter. In addition, we show in this chapter that it is possible to identify the multiple meanings of a tag by using unsupervised methods, which are more useful than consulting external resources that usually give more information than required but not adequate relevant information at the same time. We have also successfully developed a method to apply results of tag contextualisation to a real world problem—Web search result classification—and obtained satisfactory results.

The social meanings of tags discussed in this chapter is precisely one type of collective semantics that this thesis is concerned with. Users of a collaborative tagging system collectively generate implicit but meaningful associations between different

tags and documents, and the tags and documents associated with a particular tag provide the context by which its meaning can be understood.

In the next chapter, we will turn our attention to qualities of users and study how we can rank users according to their expertise in a particular topic while avoiding mistaking malicious but very active users as experts.

Chapter 5

Implicit Endorsement and Expertise Ranking

Collaborative tagging systems not only offer Web users a new method to organise their favourite Web resources, they also allow users to share what they have found interesting on the Web with other users. The collaborative nature of these systems provides the community of users a new way of discovering interesting or useful resources through other users. Tags are, obviously, very useful in helping users to retrieve relevant resources. After choosing a tag that describes resources addressing a particular topic, a user will be presented with a list of resources that have been assigned that tag. He/she can further refine this list by adding more tags to his/her query to request resources satisfying a conjunction of this set of tags. On the other hand, as the tags used by a user and his/her collection of resources are usually publicly accessible, one can also follow some users who are good at a particular topic. As these users add new resources to their collection, we can benefit from them as they act as a filter of relevant and important information.

Nevertheless, as collaborative tagging becomes more and more popular, the number of tags, users and documents involved in a system would become larger and larger. Given a list of resources that have been assigned a particular tag, it becomes desirable to have a good ranking of the resources such that we can identify high quality resources more efficiently. The same demand applies to the users as well. We need a reasonable and reliable ranking of users—ideally with respect to the topics we are interested in—such that we can identify the good users to follow. To provide a good ranking of the elements in a folksonomy is not a trivial task, and this is actually complicated by the existence of malicious users and content. As a collaborative tagging system attracts more and more users, it becomes at

the same time more attractive to spammers who would like to promote their own content. A ranking of users must therefore be able to distinguish the differences between legitimate users and spammers.

While we cannot rely on the users themselves to indicate that they are knowledgeable in a particular topic or are legitimate users who are not intended to promote irrelevant or malicious content, we can rely on the collaborative nature of tagging to determine which users are best to follow and which users (spammers) are best to avoid. We argue that the act of assigning a particular tag to a document as a previous user does can be considered as an endorsement of the previous user, although probably no users would be aware of this when they perform tagging. In this chapter, we describe in detail this notion of implicit endorsement, investigate how this idea can be used to rank the expertise of the users to facilitate resource discovery in folksonomies, and test our hypothesis regarding user expertise in collaborative tagging:

> **Hypothesis 2 (user expertise)**: The trustworthiness or expertise of the users in a collaborative tagging system can be derived from analysis of the implicit interactions among the users themselves in their tagging activities.

5.1 Resource Discovery in Folksonomies

One reason of the popularity of collaborative tagging is that users can usually discover new and interesting resources every time they visit the system, thanks to the continuous contributions of the large number of users. In existing collaborative tagging systems, there are several different ways a user can follow to discover useful resources, and these can be generally classified into two major categories: (1) following ***tags*** and (2) following ***users***.

As we have discussed in Chapter 2, tags act as indices of documents submitted by the users. One can choose a particular tag and obtain a list of documents that have been assigned the tag. In many existing collaborative tagging systems, users can also subscribe to the RSS feed of a particular tag.[1] The feed is constantly updated to present a list of documents that have been recently assigned

[1] In Delicious, the feed of, for example, the tag `photography` can be accessed at: `http://feeds.delicious.com/v2/rss/tag/photography`. In LibraryThing, the feed of, say, the tag `fiction` can be accessed at: `http://www.librarything.com/rss/tags/fiction`.

the tag. Users subscribing to these feeds can then obtain the most updated list of documents easily. The second way of discovering new resources is to follow the users. Every user has his/her own interests and therefore has his/her own collection of documents. If a user finds another user whose interests are similar to his/her, he/she can check out the collection of this user occasionally and see whether the user has added something new to his/her collection. Since this user is likely to collect resources that are related to his/her interests, these resources will also be likely to be relevant or useful to the first user. For example, in Delicious and LibraryThing, in addition to follow a particular tag, one can also subscribe to the feed of a particular user, and be notified when this user has added some new resources. These two methods are not necessarily independent of each other, and in some systems like Delicious one can also subscribe to a feed that presents resources that have been assigned a particular tag by a particular user.

While both of these two methods have their own merits and may be preferred under different situations, we believe that in general it is more beneficial to follow a user who is knowledgeable in a topic and who is more likely to add useful resources relevant to the topic. Firstly, a list of resources that have been assigned a particular tag still contains a large amount of information. In existing collaborative tagging systems, this list of resources is usually only presented according to their popularity or to how recent they have been added to some users' collection. However, neither one nor the other of these two methods of presentation guarantees that the most relevant resources are presented first. Such a list inevitably contains both high and low quality resources. On the other hand, given that we have identified a user who has the expertise in the topic in question, we can reasonably believe that resources that this user would add to his/her collection in the future will be of high quality and be relevant to the topic. Such an expert user actually acts an additional filter of the large amount of resources available in the system.

Of course, when choosing an expert user to follow, we face the question of who, of all the users who are using the collaborative tagging system, are the expert users that we should follow. To answer this question, it is necessary to first have an idea of what makes a user an expert. What are the criteria of being considered as an expert with respect to a certain topic *in the context of collaborative tagging*? And how can we measure the *level of expertise* of a user? We will discuss these questions in detail in the following section.

5.2 Expertise in Collaborative Tagging

In order to identify experts and to rank users according to their expertise, it is necessary to first have an idea of the characteristics we are looking for in an expert. In a general context, an expert is someone having a high level of knowledge, technique or skills in a particular domain. It implies that experts are individuals that we can treat as reliable sources of relevant resources and information. This general idea can be readily apply to the context of collaborative tagging. In this section, we describe and justify two assumptions we have for experts in a collaborative tagging system.

5.2.1 User Expertise and Document Quality

Given the fact that in a collaborative tagging system there are both users who are highly active and users who have only a very small collection of documents, it is tempting to judge the expertise of a user by the number of documents he/she has tagged with a certain tag representing the topic in question. For example, we may consider that a user who has tagged 100 documents with the tag `photography` is a better expert on photography than another user who has only 50 documents in his/her collection. In other words, the simplest way to assess the expertise of a user in a given topic is by the number of times he/she has used the corresponding tag (or a set of tags) on some documents. This approach is commonly used in existing collaborative tagging systems. For example, on any page that is dedicated to a particular tag, LibraryThing presents a list of the top users of that tag.[2]

However, such simple method does not consider the obvious fact that quantity does not imply quality. Knowing a lot of documents about photography does not necessarily mean that the user is an expert in the topic if these documents are not good documents that give useful and accurate information. It is also vulnerable to malicious activities such as spamming in collaborative tagging systems. As reported in some studies, highly active users are very likely to be spammers who post to the system a large amount of documents that are probably not interested by the majority of users. For example, it is found that out of the 20 most active users in Delicious, 19 of them are observed to be involved in spamming activities (Wetzker et al., 2008).

Studies in psychology explain that expertise involves the ability to select the most

[2]See, for example, `http://www.librarything.com/tag/fiction`

relevant information for achieving a particular goal (Feltovich et al., 2006). Experts also have the ability to process and apply new information faster than non-experts (Salthouse, 1991). In the context of collaborative tagging, users generally assign tags to resources so as to facilitate retrieval in case the resources are useful to their information needs in the future. A link between studies in psychology and collaborative tagging can thus be drawn. We believe that an expert should be someone who not only has a large collection of documents annotated with a particular tag, but should also be someone who tends to add *high quality* documents to their collections. The only problem here is that we now need a measure that can reflect the quality of a document. It would be quite impossible to objectively determine the quality of a document given the information we can obtain from a collaborative tagging system. However, we now know one characteristic of high quality documents given our first assumption of experts: high quality documents are more likely to be collected by expert users. In other words, the number of experts and the level of expertise of these experts who have tagged a document are good indicators of the quality of this document. In summary, there is a relationship of mutual reinforcement between the expertise of a user and the quality of a document. Expertise of a user depends on the number as well as the quality of the documents he/she has tagged, and the quality of a document depends on the number as well as the expertise of the users who have tagged it.

This approach of assessing expertise of users and quality of documents is similar to the idea behind the HITS (Hypertext Induced Topic Search) algorithm (Kleinberg, 1999) for analysing hyperlink structures on the Web. In HITS, Web documents are characterised by two properties, namely *hubness* and *authority*. The hubness of a document will be high if it has hyperlinks pointing to many different other pages (a large number of outgoing links), while the authority of a Web page will be high if it has many pages pointing to itself (a large number of incoming links). Hubness and authority have a mutually reinforcement relationship because the hubness of a page will be higher if the pages it points to have higher authority, and the same is true the other way round.

However, there is a major difference between HITS and our assumption about experts described here. In the context of collaborative tagging, there are two different kinds of interrelated entities, namely human users and Web documents, instead of only Web pages in the case of HITS. In collaborative tagging there are only links pointing from users to documents but not vice versa. This is because only users can actively assign tags to documents, while documents are passive in this setting. As a result, in our case users will only receive hub scores (expertise)

whereas documents will only receive authority scores (quality). This, however, is a very reasonable result. Experts can be considered as hubs of information because we are likely to find useful resources through them. On the other hand, high quality documents can be considered as authorities because they contain the useful information we need.

We note that this mutual reinforcement relationship and the task of co-ranking users and documents are discussed in a few studies in the literature in other contexts. For example, Zhou et al. (2007a) propose a co-ranking algorithm to rank users and scientific publications at the same time, taking into account both the social network of users and the citation network of publications. Wang et al. (2002) also describe a similar approach applied to ranking users' relevance to a certain topic by analysing hyperlinks between Weblogs. However, there are differences between collaborative tagging and the other applications discussed in the literature, as collaborative tagging offers a more general scenario. We do not necessarily have a social network of users as users are independent of each other. Also, hyperlinks between documents tagged by the users do not necessarily contribute to the evaluation of their relevance or quality, a problem we will study and discuss in detail in Chapter 6. In addition, mutual reinforcement is only one of the assumptions we have for expert users in collaborative tagging. A second and more crucial idea is what we call the notion of 'discoverer vs. follower'.

5.2.2 Discoverer vs. Follower

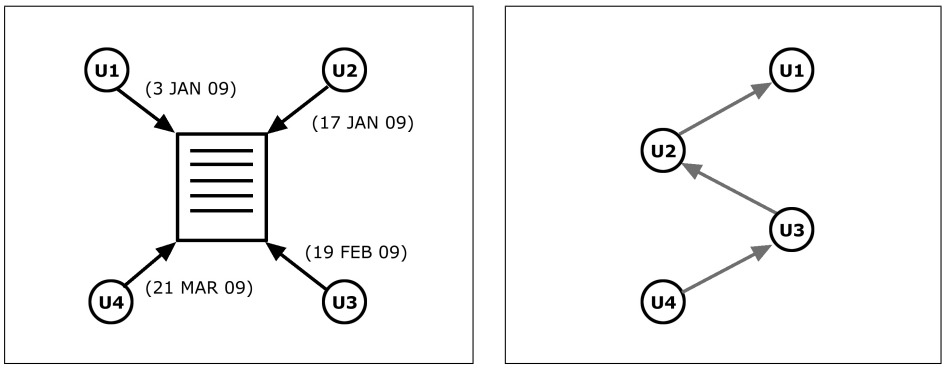
While the HITS-like mutual reinforcement approach for measuring expertise of users and quality of documents at the same time is a very intuitive and reasonable method, we still have two concerns about whether itself alone is sufficient to give good performance.

Firstly, in the HITS approach, two users will be considered to have the same level of expertise even though one is the first to tag a set of documents and the other is simply tagging the documents because they are already popular in the community. In other words, mutual reinforcement can not distinguish a user who are good at discovering high quality documents and a user who follows the examples of other users. In collaborative tagging, it is very likely that users notice new documents after some other users have tagged them and introduced them to the community. Hence, there is a great chance that users learn from each other instead of discovering information by themselves as in performing a Web search.

Secondly, such ranking method can be easily gamed by spammers. Imagine a spammer who would like to attain a high rank and attract other users to visit the irrelevant or even malicious content he/she has posted to the system, the spammer can exploit the weakness of a mutual reinforcement scheme by tagging lots of popular documents (which are likely to be of high quality) to boost his/her own expertise score. In HITS, every Web page should receive an authority score as well as a hub score. While the author of a page can increase its hub score by creating as many hyperlinks as possible to high authority pages, he/she cannot do anything alone to boost the authority score of the page because that depends whether there are hyperlinks from other pages pointing to his/her page. Hence, one can actually rely on the authority score of a page to determine its quality. However, in the context of collaborative tagging, the goodness of a user is only quantified by his/her expertise score, which is equivalent to the hub score of a page in HITS. A user can therefore actively tag a large number of documents to boost his/her own score if only the mutual reinforcement scheme is used. In other words, it is possible that spammers can manipulate their expertise scores by mimicking the behaviour of legitimate users.

Hence, in addition to knowing a lot of high quality documents per se, we believe that an expert should also be someone who is able to recognise the usefulness of a document before others do (Chi, 2006), thus becoming the first to bookmark and tag it, and by doing so bringing it to the attention of other users of the collaborative tagging system. This aspect of expertise is similar to a distinguished researcher who not only has profound knowledge of existing publications and prior art in his/her area of expertise, but who is also able to advance the field by original research of his/her own. In other words, experts should be the ***discoverers*** of high quality documents, in contrast to the ***followers*** who find these documents at a later time, because, say, the documents have already become popular or they have been featured in the mass media in the meantime. Generally speaking, the earlier a user has tagged a document, the more *credit* he/she should receive.

With this assumption, we are introducing temporal information—the *time* of tagging a document—into our analysis as an additional dimension for determining the expertise of a user. While we can never know how a user discovered a document (either by himself or by navigating within the collaborative tagging system), the time at which the user bookmarked the document is still a reasonable approximation of how sensitive he/she is to new information with respect to the topic in question.



(a) (b)

FIGURE 5.1: An illustration of implicit endorsement in collaborative tagging. (a) shows the scenario in which four users have assigned the same tag to a document at different times. (b) shows how the tagging activities can be considered as implicit endorsement between the users.

Viewing the problem from the perspective of anti-spamming measures, the notion of discoverers and followers with differing credit scores is related to protection mechanisms against Sybil attacks (Yu et al., 2006) in information security. In a Sybil attack, a malicious user creates multiple user identities in order to boost his/her reputation or 'trust score' within a system such as a peer-to-peer network. Nevertheless, an attacker can create many identities but only few trust relationships, particularly with participants outside his/her fake user network. This aspect can be exploited to identify Sybil attacks. Similarly, a spammer that floods a collaborative tagging system in order to boost his/her expertise score will end up being either just a follower (if he/she focuses on documents that are already popular within the user community) or a discoverer without any followers (if he/she introduces his/her own spam documents to the community that nobody else cares about). In both cases, he/she will not benefit much from his/her malicious activities.

The notion of discoverer vs. follower also involves a certain sense of endorsement between the users (Bonacich, 1972). Let's consider an example in which User A assigns the tag `photography` to a certain document that offers tutorials on basic techniques of photography. When another User B also assigns the same tag to the document, it implies that User B also consider the tag to be a suitable description of the document. The act of User B can be considered as an ***implicit endorsement*** of the previous tag assignment by User A. It supports User A that (1) the document can be suitably be described by the tag `photography`, and (2) User A has certain knowledge in the topic represented by the tag. As more and more users perform the same tagging act on the document, User A

can be considered as more and more trustworthy in his/her assigning the tag to the document. Figure 5.1 illustrates how the collective tagging behaviour can be viewed as implicit endorsements between the users. Conversely, if a user receives no such implicit endorsement from other users, either he/she is using a wrong tag on the document, or the document is not particularly useful or relevant to the particular tag in general.

One would suggest that a straightforward method of taking advantage of this notion of implicit endorsement to rank users is to apply the PageRank algorithm to the resultant network. However, there is a drawback of considering a network that involves only the users. For users with very few endorsements, in such a network we would not be able to tell whether they are simply non-expert users or they are spammers with malicious intent. This is because in such a network the expertise of a user depends only on the expertise of the users who implicitly endorse him, but not on the expertise of those whom he/she endorses. On the other hand, in a mutual reinforcement setting that involves both the users and documents, users who receive few endorsements still differ from each other by the quality of documents in their collection. Using documents as intermediates between users thus provide an extra dimension on which the expertise of a user can be judged.

In summary, we believe that the discoverer-follower assumption is both a reasonable and a desirable one because experts should be the ones who bring good documents to the attention of novices. In addition, this also makes our method of expertise ranking more resistant to the type of spammer mentioned above.

5.3 SPEAR: An Algorithm for Ranking Users

Based on the assumptions about experts in collaborative tagging described above, we propose *SPEAR (SPamming-resistant Expertise Analysis and Ranking)* as an algorithm to produce a ranking of users with respect to a set of one or more tags. Without loss of generality, we assume that the topic of interest is represented by a tag $t \in T$. We therefore focus on users who have used the tag t for annotations, and documents which have been assigned the tag t. The algorithm can be considered as a possible implementation of user ranking in folksonomies using our two assumptions as heuristics to assess the expertise of users.

The first step of the algorithm is to extract a set of taggings R_t from the folksonomy

$\mathcal{F}$. As we also take into consideration the time at which a tagging is created, we extend the notion of tagging by associating a timestamp to each tagging. Hence, every tagging becomes a tuple of the form: $r = (u, t, d, c)$ where c is the time when user u assigned the tag t to document d, and $c_1 < c_2$ if c_1 refers to an earlier time than c_2.

Since our algorithm is based on the HITS (Hypertext Induced Topic Search) algorithm (Kleinberg, 1999), we therefore first give a brief introduction of this algorithm before describing in detail our proposed SPEAR algorithm.

5.3.1 The HITS Algorithm

The HITS algorithm is an algorithm that performs link analysis in order to produce a ranking of Web documents. It measures two characteristics of documents, namely authority and hubness. Authoritative documents are those that provide good information with respect to a chosen topic, while hubs are documents that points to good authorities.

According to the assumptions of the algorithm, these two characteristics have a mutual reinforcement relationship: a document has high authority if many documents pointing to it have high hubness, and a document has high hubness if it points to many documents with high authority. Mathematically, the authority $a(d)$ and hubness $h(d)$ of a document d can be defined as follows:

$$a(d) \leftarrow \sum_{d' \in P(d)} h(d') \tag{5.1}$$

$$h(d) \leftarrow \sum_{d' \in C(d)} a(d') \tag{5.2}$$

where $P(d)$ is the set of documents with a link to d, and $C(d)$ is the set of documents pointed to by d.

The above operations can be represented using linear algebra. Let $\vec{a}$ be an n-dimensional vector of authority weights and $\vec{h}$ be another n-dimensional vector of hubness weights for n documents. In addition, let $\mathbf{A}$ be an $n \times n$ square matrix such that $\mathbf{A}_{i,j} = 1$ if document d_i has a link to document d_j, and $\mathbf{A}_{i,j} = 0$ otherwise. Then the algorithm at the kth iteration can be represented by the

following equations:

$$\vec{a}_k = \alpha_k \mathbf{A}^T \vec{h}_{k-1} \quad (5.3)$$

$$\vec{h}_k = \beta_k \mathbf{A} \vec{a}_{k-1} \quad (5.4)$$

where α_k and β_k are normalization constants.

The authority and hubness vectors can be proved to converge. By solving the above two equations, we have the following equations after k iterations:

$$\vec{a}_k = \theta_k (\mathbf{A}^T \mathbf{A})^{k-1} \mathbf{A}^T \mathbf{1} \quad (5.5)$$

$$\vec{h}_k = \psi_k (\mathbf{A}\mathbf{A}^T)^k \mathbf{1} \quad (5.6)$$

where θ_k and ψ_k are normalization constants. Since $(\mathbf{A}^T\mathbf{A})$ and $(\mathbf{A}\mathbf{A}^T)$ are symmetric, we can obtain for each of the matrices a set of eigenvalues with full eigenspaces. According to theories in linear algebra, $\vec{h}$ would converge to the principle eigenvector (corresponding to the largest eigenvalue) of the matrix $(\mathbf{A}\mathbf{A}^T)$, and a similar case applies to $\vec{a}$. It is found that in practise the two vectors converge quite rapidly.

5.3.2 The SPEAR Algorithm

We now describe our proposed algorithm for ranking users in a collaborative tagging system by taking into the two assumptions of experts mentioned earlier in this chapter.

Our first assumption of experts involves the level of expertise of the users and the quality of the documents mutually reinforcing each other. We define $\vec{E}$ as a vector of *expertise scores* of users: $\vec{E} = (e_1, e_2, ..., e_M)$ where $M = |U_t|$ is the number of unique users in R_t. In addition, we define $\vec{Q}$ as a vector of *quality scores* of documents: $\vec{Q} = (q_1, q_2, ..., q_N)$ where $N = |D_t|$ is the number of unique documents in R_t. $\vec{E}$ and $\vec{Q}$ are initialized by setting every element to 1. Basically, the exact value of the elements can be arbitrary as long as they are all equal, as the vectors will be normalized in later operations.

Mutual reinforcement refers to the idea that the expertise score of a user depends on the quality scores of the documents which he/she tags with t, and the quality score of a document depends on the expertise score of the users who assign tag t to it. We prepare an adjacency matrix $\mathbf{A}$ of size $M \times N$ where $\mathbf{A}_{i,j} := 1$ if user i

has assigned t to document j, and $\mathbf{A}_{i,j} := 0$ otherwise. Based on this matrix, the calculation of expertise and quality scores is an iterative process similar to that of the HITS algorithm:

$$\vec{E}_k = \alpha_k \mathbf{A}^T \vec{Q}_{k-1} \tag{5.7}$$

$$\vec{Q}_k = \beta_k \mathbf{A} \vec{E}_{k-1} \tag{5.8}$$

To implement the idea of discoverers and followers, we prepare the adjacency matrix $\mathbf{A}$ in a way different from the above method of assigning either 0 or 1 to its cells. Before the iterative process we use the following equation to populate the adjacency matrix $\mathbf{A}$:

$$\mathbf{A}_{i,j} = |\{u|(u,t,d_j,c),(u_i,t,d_j,c_i) \in R_t \wedge c_i < c\}| + 1 \tag{5.9}$$

According to Equation 5.9, the cell $\mathbf{A}_{i,j}$ is equal to 1 plus the number of users who have assigned tag t to document d_j after user u_i. Hence, if u_i is the first to assign t to d_j, $\mathbf{A}_{i,j}$ will be equal to the total number of users who have assigned t to d_j. If u_i is the most recent user to assign t to d_j, $\mathbf{A}_{i,j}$ will be equal to 1. The effect of such an initialization of matrix A is that we have a sorted timeline of any users who tagged a given document d_j.

The last step is to assign proper credit scores to users by applying a *credit scoring function* C to A:

$$\mathbf{A}_{i,j} = C(\mathbf{A}_{i,j}) \tag{5.10}$$

Regarding the choice of C, the most straightforward idea would be a linear credit score assignment such as $C(x) := x$. In this way, when the expertise scores are calculated by the iterative algorithm, users who tagged a document earlier will claim more of its quality score than those who tagged the document at a later time. One concern of such a linear credit score assignment is that the discoverers of a popular document will receive a comparatively higher expertise score even though they might have not contributed any other documents thereafter.

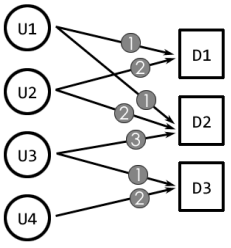
We believe that one criterion of a proper credit scoring function C is that it should be an increasing function with a decreasing first derivative: $C'(x) > 0$ and $C''(x) \leq 0$. In other words, the function should retain the ordering of the scores in A so that discoverers still score higher than followers but it should reduce the differences between scores which are too high. This is because it is undesirable to give high

Algorithm 2 SPEAR: SPamming-resistant Expertise Analysis and Ranking

Input: Number of Users M
Input: Number of Documents N
Input: A set of taggings $R_t = \{(u, t, d, c)\}$
Input: Credit scoring function C
Input: Number of iterations k
Output: A ranked list L of users.

Set $\vec{E}$ to be the vector $(1, 1, ..., 1) \in \mathbb{Q}^M$
Set $\vec{Q}$ to be the vector $(1, 1, ..., 1) \in \mathbb{Q}^N$
$A \leftarrow GenerateAdjacencyMatrix(R_t, C)$
for $i = 1$ to k **do**
    $\vec{E} \leftarrow \vec{Q} \times A^T$
    $\vec{Q} \leftarrow \vec{E} \times A$
    Normalize $\vec{E}$
    Normalize $\vec{Q}$
end for
$L \leftarrow$ Sort users by their expertise score in $\vec{E}$
return L

(a)



(b)

	D1	D2	D3
U1	1.4	1.7	0.0
U2	1.0	1.4	0.0
U3	0.0	1.0	1.4
U4	0.0	0.0	1.0

(c)

	Rank	Score
U1	1	0.422
U2	2	0.328
U3	3	0.212
U4	4	0.038

TABLE 5.1: A simple example of using SPEAR to rank users in a folksonomy. (a) A bipartite graph of users and documents. Numbers in circle represent the order of tagging by the users. (b) The adjacency matrix after credit score function is applied. (c) The final ranking of the four users.

expertise scores to users who happened to be the first few to tag a very popular document but have not contributed any other high quality documents thereafter. Here, we conduct our experiments with $C(x) := x^{0.5} = \sqrt{x}$. Overall, the above procedures of generating an adjacency matrix for the operation of SPEAR from the tagging data given a certain credit score function can be represented by the following function:

$$\mathbf{A} = GenerateAdjacencyMatrix(R_t, C) \tag{5.11}$$

The final SPEAR algorithm is shown in pseudocode in Algorithm 2, while Table 5.1 presents an example of running SPEAR on a simple case.

The SPEAR algorithm is different from the HITS algorithm in two aspects. Firstly, the adjacency matrix is not a square matrix. This is because, instead of considering a single set of documents, we now consider a set of users and a set of documents, and the number of users does not necessarily equal to the number of documents under consideration. Secondly, instead of having only 1 or 0 for the cells in the adjacency matrix $\mathbf{A}$, we initialize the matrix with different values depending on when the documents were tagged by the users. However, SPEAR can be proved to converge in the same way as HITS. This is because the proof involves the eigenvectors of the matrices $(\mathbf{A}^T\mathbf{A})$ and $(\mathbf{A}\mathbf{A}^T)$, instead of $\mathbf{A}$ (Farahat et al., 2006). Also, the proof is independent of the values in the cells of $\mathbf{A}$, as long as $\mathbf{A}$ is non-negative, which is also true in the case of SPEAR. Hence, SPEAR is guaranteed to converge under the same conditions as HITS.[3]

5.4 Experiments and Evaluation

5.4.1 Methodology

It is not a trivial task to study and evaluate the performance of SPEAR. This is because a proper ground truth of user expertise cannot be easily obtained. Firstly, there is no standard dataset for the evaluation of user ranking on Delicious. Secondly, a manual examination of the user accounts on Delicious can only be applied to a relatively small number of users, and may not result in an objective basis for the evaluation of SPEAR. To mitigate this problem, we introduce a special experimentation method that combines both real-world and simulated data to evaluate and compare the behaviour and performance of SPEAR with other baseline algorithms. We also conduct some qualitative studies to gain more insight into the performance of SPEAR.

As in our tag contextualisation studies, we rely on the data sets described in Chapter 3. To perform the evaluation, we require data sets that involve a wide range of topics, so that we can study how consistent the performance of SPEAR is across different documents and users in a folksonomy. In this sense, the data sets we have collected are actually very suitable for our purpose here as each data set contains documents that have been assigned a particular tag.

Firstly, real-world data is used as the base input for our experiments. We then

[3]In our experiments, it takes on average 160 iterations for the values in the vectors to stablize.

User Type	**Variants**	**User Type**	**Variants**
Expert	Geek	Spammer	Flooder
	Veteran		Promoter
	Newcomer		Trojan

TABLE 5.2: The simulated user profiles created for the evaluation of SPEAR.

insert controlled, simulated data into the original real-world data at the proper places. The behaviour of simulated users are determined by referring to recent studies of collaborative tagging systems (Koutrika et al., 2008; Wetzker et al., 2008) and the characteristics of our real-world data sets. By using this approach of combining real-world and simulated data, we can thus mitigate the lack of a proper ground truth by embedding controlled data into a real-world scenario, and analyse how the expected results compare to the experimental outcomes.

Regarding the simulation, to simulate a discoverer-type user, we would insert a virtual bookmark early in the timeline of a document's 'real' bookmarking history. All users with a later bookmark would automatically become followers of the simulated user for this document. Similarly, we would have to insert virtual bookmarks to popular documents in order to simulate experts because these users tend to tag only relevant information. In our experiments, we consider two major types of user profiles, namely expert-like users and spammer-like users. For each type of these users, we model three different variants in order to better match real-world scenarios and to improve the evaluation setup. An overview is shown in Table 5.2. In the following sections, we describe in detail the characteristics of each type of users and the parameters we use to simulate them.

5.4.1.1 Simulated Experts

Simulated expert profiles are subdivided into geeks, veterans, and newcomers. They represent expert users having different level of expertise with respect to a particular topic.

A *veteran* is a user who bookmarks significantly more documents than the average user, following the reports of user behaviour on Delicious described by Heymann et al. (2008a) and Noll and Meinel (2007b). He/she tends to be among the first users to tag documents which eventually become quite popular within the community. Hence, he/she is a discoverer with many followers. In the real-world, a veteran could be compared to an experienced researcher who has profound knowledge of his/her area of expertise, and advances the field by publications of his/her

own.

A *newcomer* is an upcoming expert who is only sometimes among the first to 'discover' a document. Most of the time, the documents are already quite well-known within the community at the time he/she tags them. In the real-world, a newcomer could be compared to a PhD student who already has knowledge about the state of the art in his/her area of expertise, but has yet to gain his/her reputation within the scientific community. He/she has just started with his/her own original research, so the number of publications is still low.

A *geek* is similar to a veteran but has significantly more bookmarks than a veteran. In the real-world, he/she could be a very distinguished researcher with the best knowledge of his/her area of expertise and a significant number of own publications. We can consider the geek profile as the 'best' expert within our simulation.

In the experiments, geeks should generally be ranked higher than veterans, and the latter should in turn rank higher than newcomers. It should be noted that that the differences between geeks and veterans are more subtle compared to those between veterans and newcomers. Although geeks should have a higher chance of tagging high quality documents due to their larger collections, the expertise scores of some veterans may be higher than those of some geeks. This is because SPEAR focuses on both quality and quantity instead of only quality. A user who tags a relatively small number of extremely high quality documents can still be considered as an expert.

5.4.1.2 Simulated Spammers

Simulated spammer profiles are subdivided into flooders, promoters, and trojans. They represent different types of spammers, all having malicious intentions but using different methods to boost the prominence of themselves as well as to promote their documents, which are not likely to be interested by legitimate users.

A *flooder* tags a huge number of documents which already exist in the system, most likely in an automated way. This spammer variant can often be found in existing collaborative tagging systems (Koutrika et al., 2008; Wetzker et al., 2008). He/she tends to be one of the last users in the bookmarking timeline.[4] In addition,

[4]This spammer behaviour is not only caused by specific spamming strategies which try to boost expertise/reputation scores by spamming popular documents. In practice, such behaviour can also be the result of the spam bot being created by its masters long after the Delicious service went online in 2005, so regular users have had a head start. Back in 2005, the eventual

a flooder tends to tag documents that are already known to the community, rather than tagging new documents to benefit the community. This is because his/her primary objective is to increase his/her own 'reputation' within the system by adding lots of bookmarks of existing popular content.

A *promoter* is a spammer who focuses on tagging his/her own documents to promote their popularity, and does not care much for other documents. He/she tends to be the first to bookmark documents that attract few, if any, followers. This type of spammers is actually easily observed in existing collaborative tagging systems. We find quite a number of them in Delicious while conducting our experiments. There are even cooperating groups of them who have sequentially named user accounts of the form *iSpamYou001*, *iSpamYou002*, etc., who are possibly trying to perform a Sybil-type attack as discussed in Section 5.2.2.

A *trojan* is a more sophisticated spammer in that his/her strategy is to mimic regular users in the majority of his/her tagging activities, thus sharing some traits with a so-called slow-poisoning attack. He/she disguises his/her malicious intents by tagging already popular pages, but at some point he/she adds links to his/her own documents which can be malware-infected or phishing Web pages. In other words, this spammer follows the 'majority' opinion in the folksonomy most of the time to avoid detection. He then tries to trick users into believing that he/she is a knowledgeable, benevolent member of the community and then lures them into a trap, like a wolf in sheep's clothing.

As flooders and promoters can already be observed in existing collaborative tagging systems, an algorithm for telling experts from spammers should therefore be able to handle such spammer types. Trojan-type spammers could be seen as the next step in the evolution of malicious spamming techniques, so we are interested in finding out how well SPEAR performs on these sneaky and potentially more harmful spammers.

5.4.1.3 Simulation Parameters

To simulate the different types of users described above, we manipulate the following four parameters to model their tagging behaviour.

- **P1 (Number of a user's bookmarks)**. For example, geeks and flooders

success of Delicious was not foreseeable, meaning that spamming it right away was not worth the risk and effort.

would have a greater number of bookmarks than veterans or promoters, respectively.

- **P2 (Newness)**. Percentage of bookmarks to such documents which are not in the original real-world data. To make our experiments more realistic, we needed a feature which allows simulated users to bookmark new documents, i.e. documents that haven not been bookmarked by any real-world user yet. For example, trojans and promoters create links to their own Web documents. The actual URLs of such 'new' documents are irrelevant in our experiments as long as they are unique.

- **P3 (Document rank preferences)**. A probability mass function (PMF) which specifies whether rather popular or rather unpopular documents tend to be selected when inserting simulated bookmarks. For example, the PMFs of veterans and trojans tend to select popular documents whereas the PMFs of flooders are more evenly distributed.

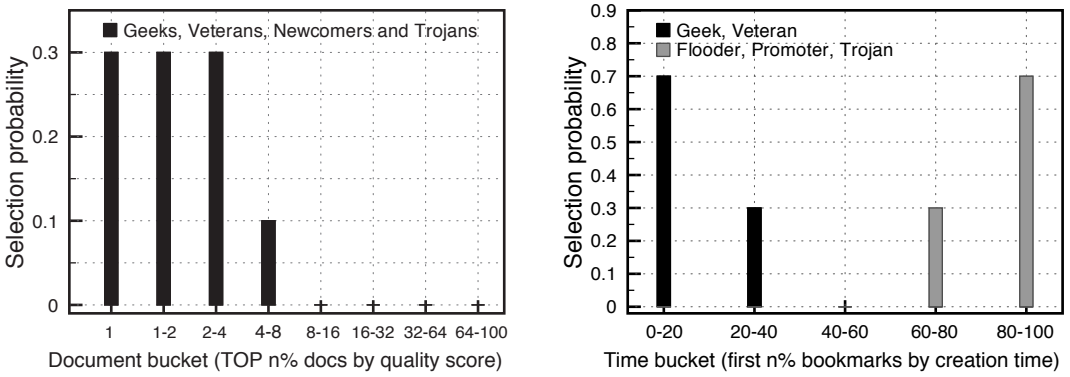
- **P4 (Time preferences)**. A probability mass function (PMF) which specifies where in the original timeline a simulated bookmark tends to be inserted into a given document's bookmarking history. For example, the PMFs of veterans tend to focus on the early stages of the bookmarking history, newcomers are rather evenly distributed, and flooders tend to be very late.

The actual configurations of the simulation parameters for each user type are shown in Table 5.3 (see also Figure 5.2(a) and 5.2(b) for the probability mass functions for **P3** and **P4**). Note that the number of bookmarks for promoters and trojans is set to absolute values (from 10 to 100), unlike that for flooders. Our reason for this decision is that promoters and trojans should exhibit behaviour similar to that of real users (flooders are more likely to be bots that generate bookmarks automatically). The mean maximum number of bookmarks of real users in our data set is $\mu_{max} = 69$, therefore our chosen values cover a similar range.

It should be noted that our simulations were probabilistic so that even identical user profiles would produce variations in simulated data. On one hand, this means that even two users generated under the same profile would behave differently up to a certain extent. For example, a 'good' geek might receive a higher expertise score than a 'bad' geek. On the other hand, we can expect overlaps in user behaviour and experimental results between different user variants. For example, a 'good' newcomer might receive a higher expertise score than a 'bad' veteran.

Type	**P1**	**P2**	**P3**	**P4**
Geek	$2 * P1_{Veteran}$	0.10	*See figure 5.2(a)*	*See figure 5.2(b)*
Veteran	$\{0.01, 0.02, ..., 0.05\} \times n_d$	0.10	*See figure 5.2(a)*	*See figure 5.2(b)*
Newcomer	$P1_{Veteran}$	0.10	*See figure 5.2(a)*	EQUAL()
Flooder	$\{0.02, 0.04, ..., 0.20\} \times n_d$	0.05	EQUAL()	*See figure 5.2(b)*
Promoter	$\{10, 20, ..., 100\}$	0.95	EQUAL()	*See figure 5.2(b)*
Trojan	$\{10, 20, ..., 100\}$	0.10	*See figure 5.2(a)*	*See figure 5.2(b)*

TABLE 5.3: Configuration of parameters P1-P4 for simulated user profiles. n_d is the total number of bookmarked documents in the relevant data set. $EQUAL()$ means that each document rank or time is selected with equal probability. The sequences of numbers in curly brackets denote multiple experiments run with varying parameters as indicated.



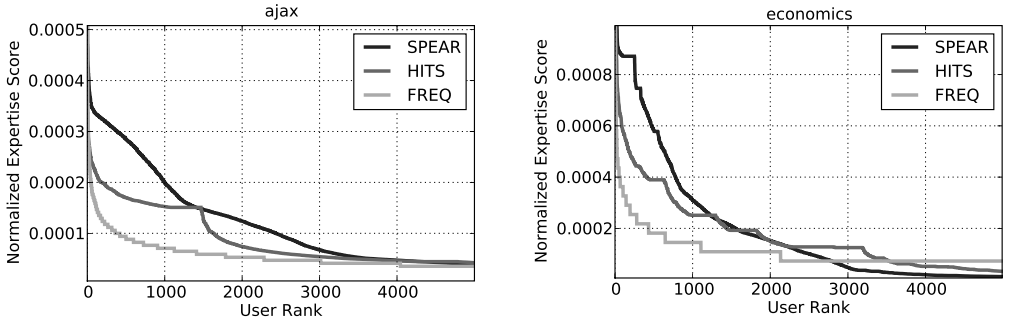
(a) PMF for document rank preferences (P3)

(b) PMF for time preferences (P4)

FIGURE 5.2: (a) PMF for document rank preferences (P3) for geeks, veterans, newcomers and trojans. Flooders and promoters chose document ranks randomly. Lower bucket numbers refer to higher quality documents. We chose exponentially increasing bucket sizes to account for the power law characteristics of folksonomies. (b) PMF for time preferences (P4) for geeks, veterans (black) and flooders, promoters, trojans (gray). Lower bucket numbers refer to earlier timestamps. In contrast to these user types, newcomers chose timestamps randomly.

5.4.2 Results and Analyses

For each of the data sets, we use SPEAR to rank both real-world users and simulated users, and compare its behaviour and performance with two other baseline algorithms. The first baseline algorithm is the original HITS algorithm. It is different from SPEAR in the initialisation of the adjacency matrix. No credit score function is applied, and each cell in the adjacency matrix either equals to 0 or 1. Hence, HITS does not consider any temporal information in the data sets. HITS would also behave similarly to other ranking algorithms for folksonomies that are derived from the original PageRank algorithm, such as the FolkRank algorithm



(a) (b)

FIGURE 5.3: Normalised expertise scores of the Top 5000 users returned by SPEAR, HITS and FREQ for two exemplary data sets: `ajax` and `economics`.

proposed by Hotho et al. (2006a) or the SocialPageRank algorithm proposed by Bao et al. (2007), as they are all based on the idea of mutual reinforcement between the entities in a folksonomy. Hence, HITS would provide us with valuable insight into how SPEAR compares with other algorithms of mutual reinforcement. The second baseline algorithm is a simple frequency count ranking algorithm, which we denote by FREQ. In FREQ, users are simply ranked by the number of times they have used the tag in question (which is same as the number of documents to which they have assigned the tag). Such simple method of ranking users is commonly found in existing collaborative tagging systems.

In the following sections, we first describe and compare the general behaviour of the three algorithms. We will then describe in detail how different types of expert users and spammers are ranked by these algorithms. We will also discuss qualitative some users that are ranked at the top and the bottom of the lists to gain more insight into the characteristics of SPEAR.

5.4.2.1 General Behaviour

Figure 5.3 shows the normalised expertise score distributions of SPEAR, HITS and FREQ for two exemplary data sets, namely `ajax` and `economics`. We observed that SPEAR generally produced more differentiated values than HITS and FREQ for top users, i.e. the difference in expertise scores between two ranks for SPEAR was generally larger than for HITS and FREQ, where the curves were flatter. We will see how SPEAR benefits from this characteristic in Section 5.4.2.3.

Another observation is the staircase-like shape of FREQ caused by the integer frequency counts on which it is based. This means that FREQ tends to group

users into buckets of equal expertise score instead of assigning an individual rank to each user. In other words, using such simple frequency counting technique to rank users will result in many users having the same rank. While such 'staircase steps' can also be observed in the cases of HITS and SPEAR, they are due to different reasons.

In HITS, two users who have assigned the same tag to the same set of documents will be assigned the same rank. This is because their expertise scores come from the quality scores of the documents, and if they collect the same set of documents they will receive the same expertise score. We see that there are actually quite a lot of these cases. For example, there are a group of users who all have tagged the same small set of hugely popular documents, resulting in their being assigned the same expertise score in HITS.

The 'staircase steps' observed in SPEAR, however, are mainly due to a limitation of our data sets. As we can only retrieve from Delicious the date instead of the exact time when a bookmark was created by a user, there are quite a lot of time collisions in the data sets, i.e. there are some users who assigned the same tag to the same document on the same day. Therefore if two users happen to have bookmarked the same set of documents and each document was bookmarked on the same day, they will receive the same expertise score. Nevertheless, this problem can be solved if we have access to a more fine-grained timeline of the documents.

In summary, SPEAR is able to spread the expertise scores of the users across a wider range, and it is much less likely to assign the same score to two users than HITS and FREQ.

5.4.2.2 Promoting Experts

To study how different variants of experts are ranked by SPEAR, we generate, for each of the real-world data sets, 20 experts of each type—a total of 60 simulated experts per data set—and insert them together with their simulated bookmarks into the corresponding data set. We then apply SPEAR, HITS and FREQ to these data sets comprising both real-world and simulated users. We present the results in two different ways.

Firstly, we normalise the rank of the simulated users using the following equation:

$$\text{Normalised Rank}(u) = \frac{|U_t| - \text{Rank}(u)}{|U_t|} \tag{5.12}$$

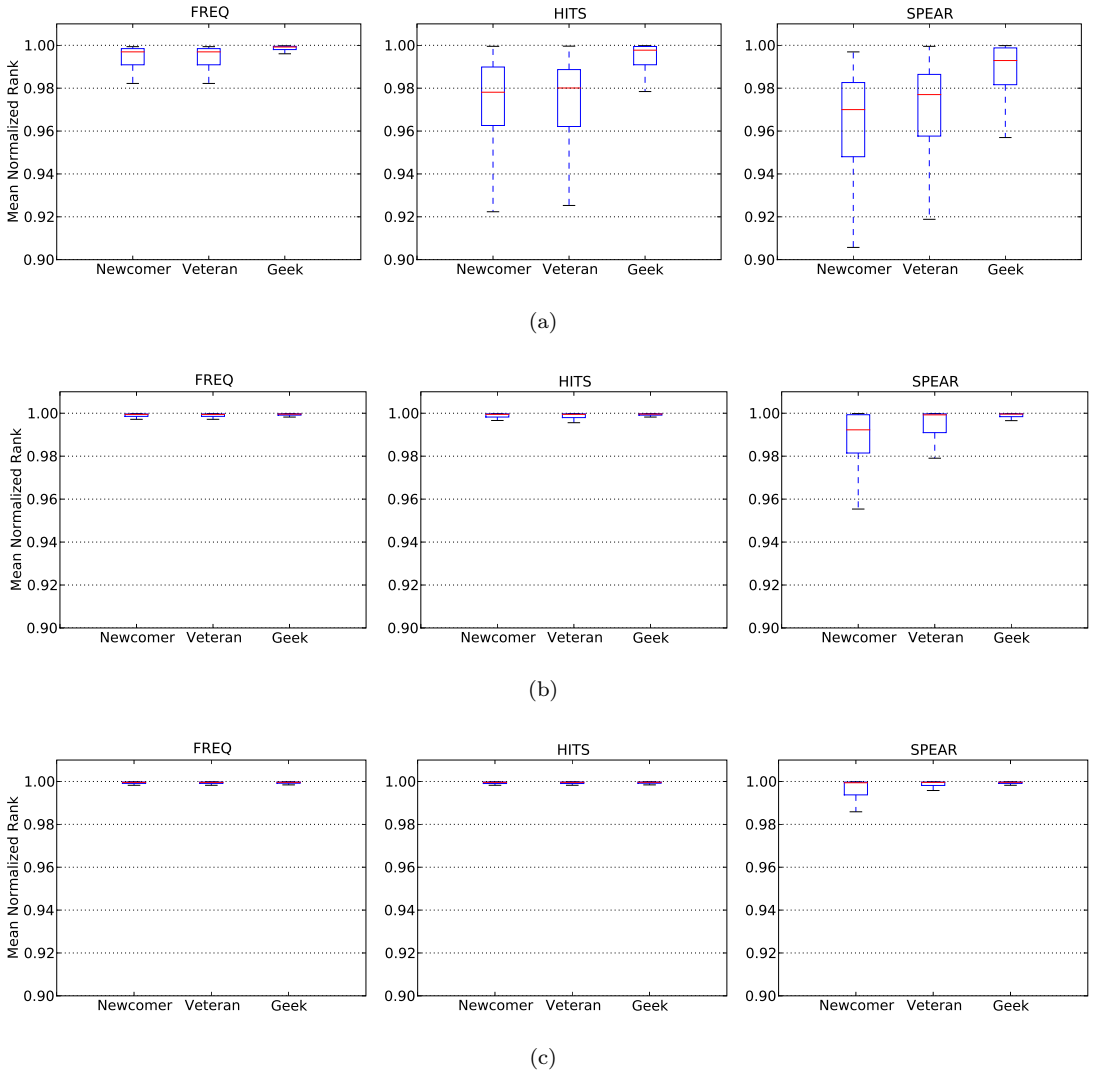
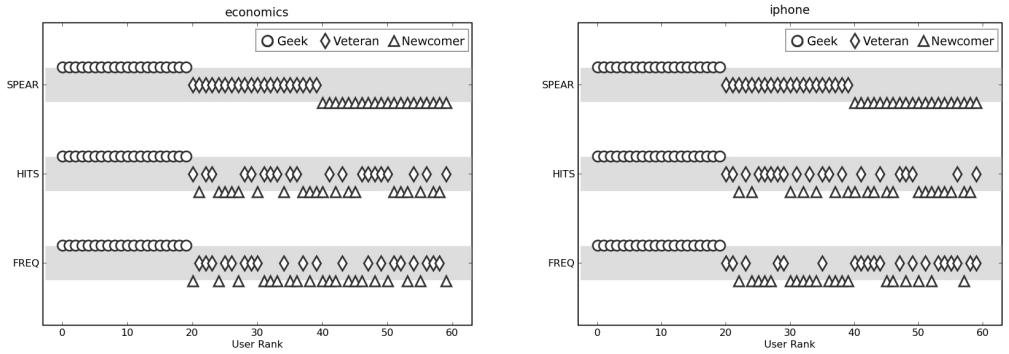


FIGURE 5.4: Boxplots of mean normalised ranks of simulated experts—**newcomers, veterans, geeks**—in direct comparison across all data sets for the three algorithms. The plots (a), (b) and (c) show the results for $P1_{Veteran} = 0.01$, $P1_{Veteran} = 0.03$ and $P1_{Veteran} = 0.05$, respectively.

In this way, a user who is assigned the highest expertise score (with rank 0 as the highest rank) will receive a normalised rank of 1.0. If on the other hand a user is assigned the lowest expertise score, he/she will receive a normalised rank of 0.0. Using this equation, we calculate the average normalised rank of the simulated experts and plot the results in Figure 5.4. It should be noted that some overlap between the three expert variants are expected due to the PMF-based simulation setup as described in Section 5.4.1.

The plots show some major differences between SPEAR and the other two ranking algorithms. In SPEAR, geeks are generally ranked higher than veterans, which in turn are ranked higher than newcomers. We also observe that geeks and experts



(a) Simulated experts in `economics`

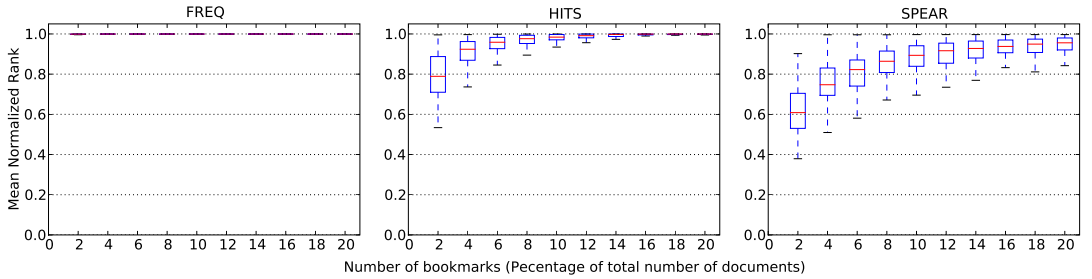
(b) Simulated experts in `iphone`

FIGURE 5.5: Ranks of simulated experts for two selected tags `economics` and `iphone`. In (a) and (b), SPEAR clearly distinguishes between the three types of expert users, while HITS and FREQ tend to mix up veterans and newcomers.

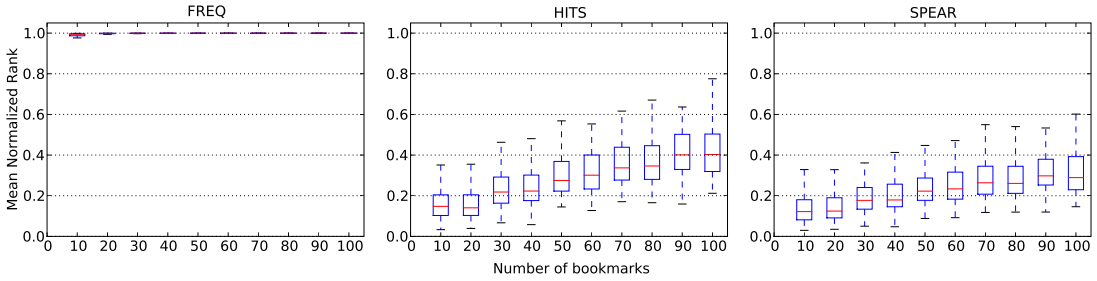
do compete for the top ranks even though geeks win in general. This means that some veterans, although having had fewer documents than geeks in general, are ranked higher by SPEAR because they have some documents of higher quality. Another observation is that veterans are ranked higher than newcomers. Here again we see some newcomers are assigned higher ranks than some veterans, due to a similar reason.

On the other hand, HITS and FREQ do not perform as good as SPEAR. While both algorithms rank geeks higher than veterans and newcomers, but geeks are also the "easiest" of all expert variants to be ranked correctly, because they have a very large number of supposedly high quality documents. This means that even the naive FREQ should and do perform reasonably for this user variant. However, both HITS and FREQ fail to differentiate veterans from newcomers, which end up being mixed with each other. This result suggests that only SPEAR succeeded in distinguishing veterans and newcomers by implementing the notion of discoverers and followers. In contrast, HITS still tend to return results that are heavily influenced and biased by the number of documents in a user's collection, even though it is also an implementation of a mutual reinforcement scheme.

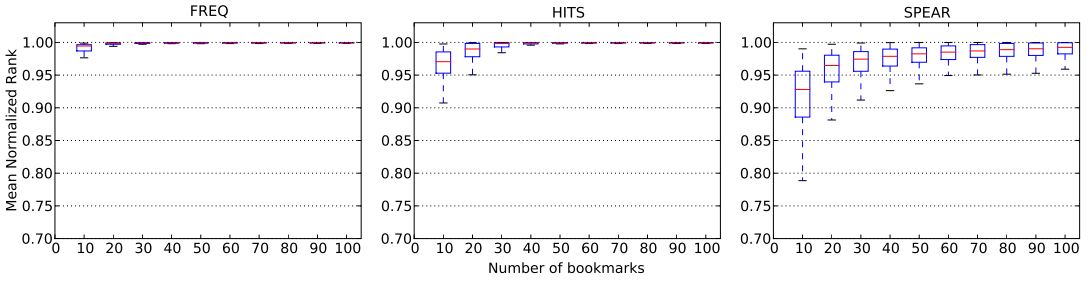
To have a closer look the differences between SPEAR and the other two algorithms, we select two tags, namely `economics` and `iphone` from our data sets and visualise the ranks of the simulated expert users as shown in Figure 5.5. We can see that the three expert variants are clearly separated by SPEAR, but veterans and newcomers are mixed up with each other in HITS and FREQ. From these results, we can conclude that in usage scenarios where quantity does not guarantee quality—and we believe collaborative tagging is one such scenario—SPEAR is expected to



(a) Flooders



(b) Promoters



(c) Trojans

FIGURE 5.6: Boxplots of mean normalised ranks of simulated spammers in relation to the number of bookmarks generated per spammer.

produce better rankings of users as it is able to detect the more subtle differences between different types of users.

5.4.2.3 Demoting Spammers

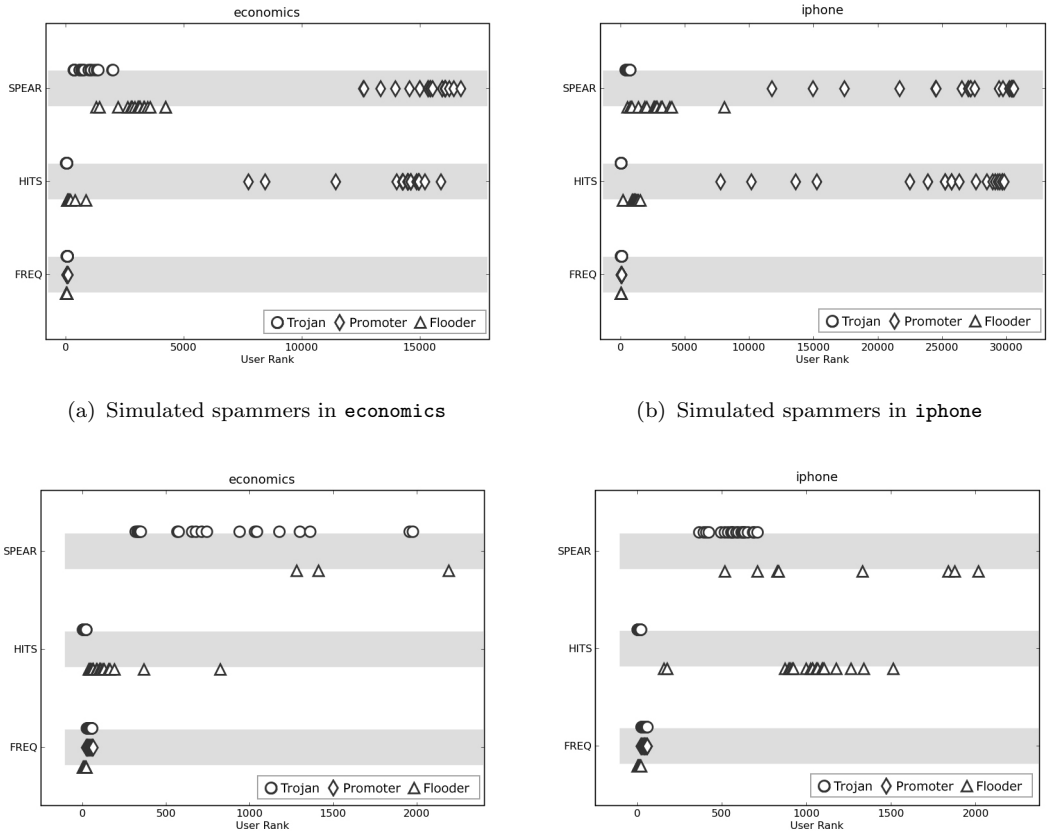
In a similar fashion as the experiments on experts, we generate 20 flooders, promoters and trojans and insert them into the real-world data sets. We also test the performance of the three different algorithms by manipulating the number of documents each spammer would tag. Again, using Equation 5.12, we calculate the average normalised rank of the simulated spammers and plot the results in Figure 5.6.

From the figures, we can easily see that FREQ shows the worst performance among the three algorithms. All spammers are ranked top users simply because they have tagged a large number of documents. This shows that a simple ordering by frequency is very vulnerable to spamming activities in a collaborative tagging system. This is particularly true for flooder-type spammers, which unfortunately are often found in today's collaborative tagging systems (Wetzker et al., 2008). HITS, on the other hand, performs better than FREQ but is dominated in all experiments by SPEAR. While HITS is good at demoting promoters, it has problems to demote flooders with increasing numbers of spam bookmarks, and is weak in general in handling trojans.

SPEAR shows the best performance among the three algorithms. Firstly, it correctly demotes both flooders and promoters by assigning them significantly lower ranks than HITS and FREQ. One may suggest that flooders are actually not as malicious as other types of spammers considered in this study, because their primary intend is to boost their own expertise score by bookmarking a lot of documents, instead of introducing malicious content. However, mistaking some users who randomly bookmark a large number of documents as expert users is clearly something that should be avoided. And in practice it is also possible that these users may introduce whatever content they like after they have attained a certain expertise score. Therefore the fact that SPEAR demotes these users can be considered as a merit of the algorithm.

In addition, SPEAR is also able to demote trojans, which use a much more sophisticated spamming scheme. While trojans are still ranked higher than the other two spammer variants, a closer look at the rankings produced by SPEAR reveals that trojans are rarely ranked higher than rank #100 across the data sets in the experiments. This characteristic of the results can be better observed in Figure 5.7, which shows the positions of the spammers for the two selected tags `economics` and `iphone`, with P1 set to $0.2 \times n_d$ for flooders, and 100 for promoters and trojans. We can see that all the simulated trojan spammers are ranked lower than the 200 by SPEAR. However, these spammers still find their way to the top ranks in HITS and FREQ. Given that in practice the TOP 10 to the TOP 50 experts should be the ones we are most interested in, SPEAR in its current form already performs reasonably well in getting rid of all trojans in the relevant rank range.

Nevertheless, we have to be aware of the problem with trojans, which is that it is tricky to demote them without demoting good users at the same time. This is because from a pragmatic point of view a trojan is still a rather good hub of



(a) Simulated spammers in `economics`

(b) Simulated spammers in `iphone`

(c) Simulated spammers in `economics` (zoomed)

(d) Simulated spammers in `iphone` (zoomed)

FIGURE 5.7: Ranks of simulated spammers for two selected tags `economics` and `iphone`. (a) and (b) show the whole range of user ranks, while (c) and (d) zoom into the top 2500 users. It can be seen that SPEAR is able to demote trojans such that they are not ranked among the top 200.

resources, given the popular documents they have tagged in the hope of boosting their own expertise scores. To avoid falling into the trap of this type of spammers, one may need to verify the quality score of the documents of highly ranked users, which is also computed by SPEAR, to judge whether they are really legitimate and useful resources before actually accessing them.

Finally, we also observe that SPEAR is the only algorithm that does not tend to 'clump' spammers together in one spot in our experiments, i.e. it is better at differentiating and detecting nuances in spammer behaviour compared to HITS and FREQ. This is probably a direct result of the different expertise score curves as described in Section 5.4.2.1.

5.4.2.4 Qualitative Analysis

In addition to the quantitative analysis of the simulation results, it is worthwhile to take a look at the ranking produced by SPEAR in a qualitative way so as to gain more insight into its effectiveness.

We run SPEAR on the data sets of four arbitrarily selected tags, namely `photography`, `semanticweb`, `javascript` and `programming`, where the last two are combined to form a conjunction as an example of running SPEAR on a more specific topic. We examine the top users who are given high ranks by SPEAR in each of these data sets. While we are likely to be able to provide an objective evaluation of the expertise of these users, we discover that there are several things that are indicative of their expertise. Firstly, many of these top users are more likely to provide optional personal information in their Delicious account, including for example their real names, address of personal Web sites, links to their photos on Flickr, and links to their Twitter microblogging account. This implies that they are more involved in using Delicious. Secondly, many of them have a lot of other tags used together with the corresponding tag in which they attain high expertise score. For example, a top user in `photography` has used 359 other tags together with `photography`, suggesting that he/she has an extensive collection of documents about the topic. Finally, we identify some 'real' experts among the top users. For example, two users that have been ranked in the top 10 in `semanticweb` turn out to be two researchers of the Semantic Web, while another is an active blogger of the same subject. The top two experts ranked by SPEAR in `javascript`$\cap$`programming` are two professional software developers. In contrast, all the users mentioned above are ranked lower by FREQ and HITS, sometimes even outside the top 200.

As for spammers, we single out the obviously heavily spammed tag in Delicious, `mortgage`, collect the bookmarking histories of the documents that have been assigned the tag (the data set of this tag is not among the 130 data sets we have collected at the beginning), and run SPEAR, HITS and FREQ on it to rank the users. We want to find out whether spammers are really demoted by SPEAR and whether FREQ is vulnerable to spammers in this real setting. While we do not have a list of the spammers, we identify them by looking for several characteristics common to spammers. Spammers are usually automated bots. Hence, they either tend to use extract words from the documents themselves (especially the title) and use them as tags, or use the same set of tags on a large number of documents. Also, some spammers aim at promoting their own content, and therefore many of their bookmarks are likely to be documents from the same domain (which can

usually be understood to be spams at first glance).

By looking for these characteristics in the users who have used the tag `mortgage`, we successfully identify 30 spammers in the top 50 most active users. Obviously, this means that out of the top 50 users ranked by FREQ, 30 of them are found to be spammers. It is interesting that we even discover a group of spammers whose usernames have the same prefix and are only different from each other in the numbers in the suffixes, suggesting that there do exist spammers who submit spams in a more sophisticated way than merely flooding the system. As for the rankings produced by SPEAR and HITS, we observe similar results as we do in our simulations. All these 30 spammers are significantly demoted to below the 3000th rank by SPEAR and HITS, with ranks of these spammers in SPEAR much lower than those in HITS. We also see that there are no spammers in the top 50 ranks returned by SPEAR and HITS.

In addition, we also run FREQ and SPEAR on arbitrarily selected tags and examine the differences between the top rank users. We find that very often users ranked at the top by FREQ are quite the opposite of experts, not to mention that many of them are spammers. For example, for the tag `bridge`, a user is ranked first by FREQ because he/she has a large number of bookmarks with the tag. However, a closer look at his/her collection of documents in Delicious reveals that the majority of them are not related to any conventional meanings of the word 'bridge', including meanings found in WordNet and those discovered in the tag contextualisation process described in Chapter 4. In contrasts, SPEAR ranks this user much lower, at 2,088th out of the 3,144 users being ranked. The fact that this user is ranked low by SPEAR is that, despite the number of times he/she has used this tag, there are very few, if any, other users who would do the same thing as he/she has done. In other words, although he/she is not a spammer, this user receives very few endorsements due to his/her idiosyncratic use of the tag. Arguably SPEAR gives a more sensible result because other users are quite unlikely to benefit from this user with respect to the topic in question.

By this small qualitative study, we show that SPEAR also works reasonably well in a real setting. On the one hand, it is able to identify real experts. On the other hand, it is able to solve real problems by demoting real spammers found in Delicious.

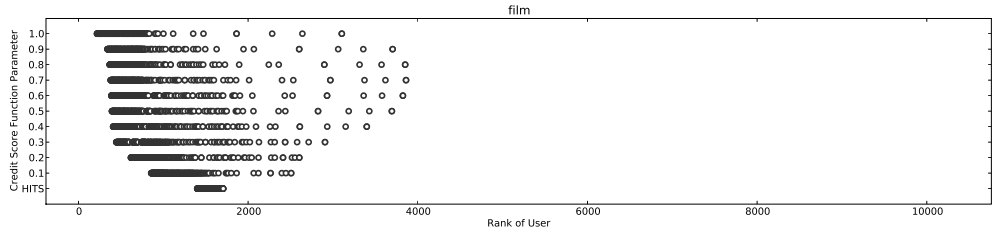
5.4.3 Analysis of Credit Score Functions

One important element of SPEAR is the credit score function $C(x)$ by which we assign higher scores to users who have tagged a document earlier and lower scores to users who have tagged the document at a later time. This credit score function actually directly affects the performance of SPEAR. If we do not apply the credit score function, SPEAR will be no different from the original HITS algorithm, in which every cell in the adjacency matrix will either be 1 or 0.

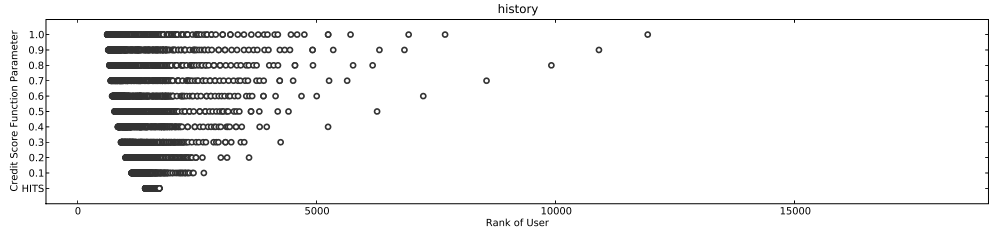
Intuitively, with a credit function of larger second derivative—credit scores for a user increases faster and faster when he/she has more and more followers, SPEAR should be more resistant to spammers. This is because the number of followers of a user is an important piece of information that allows us to distinguish between spammers from legitimate users. However, there is also a drawback when such an aggressive credit score function is used, as we have briefly mentioned in Section 5.3.

To give higher scores to users who have tagged a document at an earlier time will increase the chance of mistaking an inactive user as an expert. Consider a very popular document with 5,000 users, a certain user may happen to be the 100th user to tag this document, and therefore he/she has 4,900 followers with respect to this document. As a result, he/she will be assigned a an initial score of $x = 4,900$. Consider two credit score functions $C_1(x) = x^{0.2}$ and $C_2(x) = x^{0.8}$: $C_1(x)$ will return 5.47, while $C_2(x)$ will return 895.69. If C_2 is used, this user will receive an exceedingly high expertise score given this high credit score coupled with the probably very high quality scores of this popular document. Other expert users who have tagged many more high quality documents will find themselves ranked lower than this user only because they are followers of him in this particular document. This will be a problem because this inactive user is very unlikely to benefit other users.

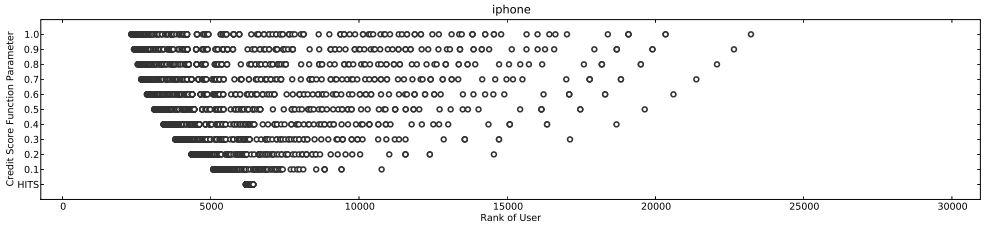
To investigate how the credit score function affects the ranks of these inactive users, we conduct some experiments on some selected data sets with different credit score functions. Firstly, we randomly pick three tags from our data sets: `film`, `history` and `iphone`. For each of these data sets, we run SPEAR to obtain a ranking of the users involved by using different credit score functions of the form $C(x) = x^y$, where y ranges from 0 to 1.0 (in the case of y=0, the algorithm effectively becomes the same as HITS). While it is true that there are many other types of functions that can be considered here, this class of functions should be sufficient in allowing us to have a better understanding of the behaviour of SPEAR, as it provides us



(a)



(b)



(c)

FIGURE 5.8: Ranks of users who have only tagged the most popular document for each of the three selected tags: `film`, `history` and `iphone`. Only these users are represented by the circular symbols. Other users in the data sets are not shown.

with functions with different second derivatives, which we are most interested in. We then examine for each of the tags the ranks of the users who are found to have only tagged the most popular document in the respective data set.

Figure 5.8 shows the ranks of users who have only tagged the most popular document in each of the three data sets, with SPEAR operating under different settings of credit score function. We can see that the differences between credit score functions show similar effects on the ranking of these inactive users. Credit score functions with greater values of y tend to spread the users across a wider range. This is due to the fact that these credit score functions assign scores that spread a wider range of values. However, these functions also tend to rank some inactive users quite high, especially when they tagged the most popular document at an very early time.

On the other hand, credit score functions with smaller values of y tend to clump

users in small range of ranks. At the extreme end where $y = 0$, all of the users under consideration are assigned the same expertise score. A merit of these functions is that they tend to give lower range to these users on average. Therefore they also have a smaller chance of mistaking these users as expert users. However, as we have shown in our simulations described in Section 5.4.2.3, HITS, which is SPEAR with $y = 0$, performs relatively poorer than SPEAR where we set $y = 0.5$. In other words, smaller values of $y = 0$ would also make SPEAR more vulnerable to spammers.

Different credit score functions have different merits and weaknesses. Therefore there is no single correct choice of credit score function for SPEAR. In settings where spamming activities are commonly observed, functions with greater values of y or other functions with similar characteristics should be used. On the other hand, in settings where there are few spammers, one may consider to use functions with smaller values of y or other functions with similar characteristics.

5.5 Discussion

Our experiments show that the algorithm we have described in this chapter, SPEAR (Spamming-resistant Expertise Analysis and Ranking), produces better rankings than both the original HITS algorithm and simple frequency counting. It is able to distinguish reasonably well between different types of experts, and it consistently demotes different types of spammers and removed them from the top of the rankings. In other words, SPEAR is able to detect the subtle differences between good and bad users, and to demote spammers while still keeping the experts at the top of the ranking. We note that SPEAR measures expertise mainly based on a user's ability to discover (new) high quality content, which is but one aspect of an expert's skill set in the real world. However, since a primary goal of collaborative tagging systems is to identify high-quality resources, the expertise aspect analysed by SPEAR is therefore very relevant in these systems.

There are a number of reasons of why an expert ranking algorithm is needed in collaborative tagging. Firstly, with increasing number of documents for a given tag, it becomes increasingly difficult to retrieve documents which are useful and of good quality. One way to solve this problem is to first identify the experts and then browse their collection which should contain good documents. On the other hand, by keeping an eye on the collection of an expert, we are able to benefit from

notification when he/she adds new and useful documents to his/her collection.[5]

The promising results achieved by SPEAR and the relatively poor results achieved by HITS and FREQ in terms of their resistant to spamming activities suggest that there is a risk in directly borrowing recursive algorithms from link structure analysis on the Web to ranking entities in a folksonomy. This is due to a fundamental difference between hyperlinks on the Web and relations between users in a folksonomy. An author of a Web document can create as many hyperlinks to other documents as he/she wants, but does not have control on which documents would link to his/her own. In other words, he/she cannot manipulate the authority score of the document in HITS or the its PageRank score easily.[6] On the other hand, in a folksonomy all we can depend on in general to assess the expertise of the users are their collections of bookmarks and tags, both of which can easily be expanded or modified by the users. Directly applying recursive algorithms to assess their expertise will result in rankings that are heavily influenced by the activeness of the users. The incorporation of the notion of implicit endorsement (and therefore temporal information) by SPEAR is an effective solution to this problem because the time of tagging is managed by Delicious (a trusted entity) and cannot be manipulated by the users.

While our discussion centres on expertise and experts, SPEAR can also be considered as an algorithm for assessing the trustworthiness or the reputation of the users (Resnick et al., 2000; Resnick and Zeckhauser, 2002). The number of followers a user has or the number of implicit endorsement a user has received can be considered as an indication of how trustworthy the user is. The score can also be considered as how reputable a user is in tagging useful documents with respect to a particular topic. Nevertheless, as we have discussed at the beginning of this chapter, collaborative tagging is more about resource discovery and therefore it is reasonable to focus on the users' knowledge and ability to introduce useful and relevant resources to other users.

Our study described in this chapter demonstrates that some important features of the users themselves can actually be revealed by analysing the collective behaviour of the users participating in collaborative tagging. The notion of implicit

[5]Currently, Delicious allows users to subscribe to a particular tag or be a fan of an other user. However, there is neither a measure of a user's expertise nor a recommendation of related experts in your areas of interest given your own user profile.

[6]Of course, manipulating a document's PageRank (Becchetti et al., 2008) is possible such as by purchasing hyperlinks from highly-ranked documents. Here, we refer to the fact that an author cannot manipulate the PageRank of his/her document easily by only modifying the content and its out-going links.

endorsement is very important for understanding the expertise or trustworthiness of a user. It also shows that the solution of such 'social vandalism' as spamming in collaborative tagging can in fact be alleviated by the community of users themselves. What is required here is thus a proper way of harnessing the collective intelligence of the users.

Although we only discuss expert ranking in the context of collaborative tagging, SPEAR is in fact applicable in many different situations because it assumes a very general model of user-document interactions. For example, it can be applied to collaborative filtering sites such as Last.fm (a social Web application for sharing musical content) and Digg (a social news Web site), which are very popular among Web users nowadays, to rank users by their expertise in a given topic.

5.6 Chapter Summary

This chapter has presented our study of the notion of implicit endorsement in collaborative tagging systems. Based on our two assumptions of expert users in collaborative tagging, we have proposed the SPEAR algorithm, which takes into account the mutual reinforcement relationship between users and documents as well as temporal information of tagging, for ranking users according to their expertise in a particular topic. The algorithm has been shown to be able to promote expert users and demote spammers at the same time effectively. We have shown that by using a suitable method to aggregate the collective behaviour of the users in a collaborative tagging system, we are able to acquire a better understanding of the characteristics of the users. Here, the collective semantics we are concerned with is the trustworthiness/expertise of the users. While this information is not explicitly provided anywhere in the system, we have shown that it can be successfully derived by using SPEAR to analyse the implicit interactions between the users, thus supporting our hypothesis regarding user expertise in collaborative tagging.

In the next chapter, we will turn to study semantics of documents in a folksonomy, in particular the relations between them and how we can identify implicit relations resulted from tagging activities of users.

Chapter 6

User-induced Hyperlinks

Hyperlinks are probably the most important elements on the Web. Their existence is the reason why the Web is a web: they allow Web users to jump from one hypertext document to another, making navigation through the Web possible. Hyperlinks are generally embedded in hypertext documents. This means that very often only the author of a hypertext document can decide on which other documents this one can link to. Of course, there are personalised portal sites that generate dynamic content based on, for example, the preferences or browsing habits of users. However, the majority of hyperlinks are created from the perspective of the authors of the documents. While such perspective may be necessary when hyperlinks are created for navigation within a particular Web site, such author-created hyperlinks can be limited when they are intended to direct Web users to relevant or potentially interesting documents. Henzinger (2005) mentions two types of hyperlinks, those for navigation and those for recommendation. Recommendation hyperlinks point users to other documents that contain information related or complementary to the current document. Obviously it is possible that an author cannot always ensure his/her document has hyperlinks pointing to all the other relevant documents.

When hyperlinks in a document do not provide enough useful references to other relevant documents, users rely on other methods to seek help. One alternative is to rely on Web directories such as the Yahoo! Directory and the Open Directory, in which users can pick a category and browse the documents in the category. Another alternative is to search for relevant documents by submitting a query to search engines such as Google or Ask.com. Since the rise of popularity of Social Web applications, collaborative tagging systems have provided another alternative starting point for Web users to look for relevant information. As we have discussed

in the previous chapter, users can browse through documents that are assigned the same set of tags, or they can browse the collections of users who are found to be knowledgeable in a particular topic. Viewing this merit of collaborative tagging from a different perspective, two documents that are more relevant to each other should be more likely to 'co-occur', i.e. to be assigned the same tag or a similar set of tags by many users. We then have here an interesting question regarding collaborative tagging: are there hyperlinks between documents that have been tagged by the same group of users or have been assigned similar tags? If there exist such hyperlinks, it would suggest that the hyperlinks do help users to retrieve relevant documents by functioning as recommendations. If, however, there does not exist any hyperlink between documents that are interested by many users, it would suggest that the existing link structure between the documents are very inadequate. In other words, does the collective behaviour of the users give rise to relations between documents that were not explicit before?

Such research question actually leads to a more general question that is often posed to the Social Web. The Social Web claims to be a Web for ordinary users. Content on the Social Web is generated by users and represent a collective viewpoint of the users. In contrast, the hyperlink structure of the Web represents the viewpoints of the authors of the documents. It represents how the documents should be interconnected from the perspectives of those who provide information on the Web. So how do implicit links between documents on the Social Web different from those explicit hyperlinks created by the authors of the 'original' Web? In other words, how does the perspective of authors compare with that of readers? Some studies already point out some differences between these two perspectives. For example, Bao et al. (2007) adapts the PageRank algorithm and propose the SocialPageRank algorithm to rank documents in a folksonomy. They find out that some documents have high SocialPageRank but low PageRank, and some documents have low SocialPageRank but high PageRank. Noll and Meinel (2007b) compare tags with metadata of documents provided by the authors themselves, and reveal that tags contain additional information about the documents that cannot be found in the metadata provided by the authors. Nevertheless, we have not seen any systematic study of the differences between the authors and readers on the Web at the level of hyperlinks, which are basic building blocks of the Web.

In this chapter, we discuss our study devoted to investigating the research questions mentioned above. In particular, we test the following hypothesis in this chapter:

> **Hypothesis 3 (relations between documents)**: The implicit relations between documents generated by the collective tagging activities of Web users represent better recommendation links than existing hyperlinks created by the authors of these documents.

We employ the data mining technique of association rule mining and also other similarity measures to discover implicit hyperlinks, or what we call user-induced hyperlinks, in a folksonomy. We also compare and contrast these user-induced hyperlinks with existing hyperlinks among the documents. We aim at uncovering the differences between the perspectives of the authors and the readers on the Web, which have been made more prominent by the popularity of collaborative tagging. Finally, we also propose to predict tags for a particular document by considering other documents that are connected to it by user-induced links, and show that these links are also useful in the context of document classification.

6.1 Hyperlinks on the Web

Hyperlinks originate from the idea of linking documents to each other, or enabling cross-referencing, to facilitate reading of a large set of documents. This idea is first described in a systematic way by Bush (1945) who proposes a machine called Memex that would enable a user to create links between any pages of the books in a library in the form of microfilms. Nelson (1993) brings this concept further and introduces the Xanadu Project, which attempts to implement a hypertext system in which links can be created between any documents stored in a computer. Hyperlinks are indispensable elements in the World Wide Web, which has been the most successful implementation of hypertext systems in terms of its widespread use and popularity. One of the features of the Web that attracts so many users from all over the world is that users can follow hyperlinks embedded in a document to discover other related and useful documents. In some cases, hyperlinks help users to navigate on the Web; in other cases, they provide recommendations to users as to which other documents they can visit to obtain more information.

However, relations between documents are not limited to the explicit hyperlinks defined within the documents themselves. As a huge number of users search for information they need, browse a set of pages that they find useful, or assign tags to a set of resources that they find interesting, we would know that some documents are somehow related to each other from the perspectives of the users. Two

(a)

```
...
Document A.html
<a href="http://www.example.com/B.html">Link</a>
to B.html
...
```

```
...
Document B.html
<a href="http://www.example.com/C.html">Link</a>
to C.html
...
```

(b)

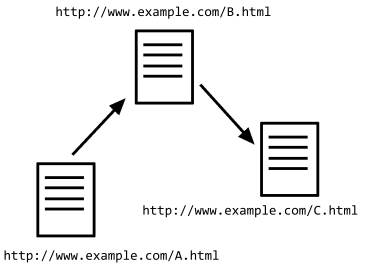


TABLE 6.1: (a) The HTML source codes of hypertext documents A.html and B.html. (b) The resultant link graph of the three documents.

documents can be deemed as related to each other if a lot of users think that both are interesting or useful to them, even though they are not connected to each other by a hyperlink. In this case, it can be said that there is an implicit link between them. Given that user preferences are collected in many different situations nowadays (such as search engine query logs or Web logs), it is not difficult to discover implicit links between documents.

In the following sections, we present a brief introduction of hyperlinks and the explicit link structure of the Web. In addition, we also discuss some previous studies that investigate various methods for generating implicit links between Web documents and how these links facilitate analysis of the relations between the documents.

6.1.1 Hyperlinks and Link Structure of the Web

A Hyperlink, or simply a link, from one Web document to another Web document is created mainly by using the standard syntax of HTML.[1] This is done by specifying the target of the link by providing the URL in the anchor tag of HTML. Hyperlinks are usually underlined in a Web document, although nowadays this feature can be easily changed by using techniques such as Cascade Style Sheets (CSS) to accommodate different designs.[2] Table 6.1 presents a simple example of two hyperlinks between three documents. Note that hyperlinks on the Web are directional and one does not necessarily know from which documents hyperlinks are originated.

The functions of a hyperlink can be roughly classified into two categories (Hen-

[1]HyperText Markup Language (HTML): `http://www.w3.org/MarkUp/`
[2]Cascade Style Sheets: `http://www.w3.org/Style/CSS/`

zinger, 2005). Firstly, hyperlinks can be used to assist navigation of Web users within a certain Web site. For example, an online catalogue of products of a company may contain a lot of hyperlinks that allow users to browse the details of the products efficiently. This kind of links can also be used by the author of a Web site to create a certain user experience, such as guiding the user through a series of Web pages that should be read sequentially. Navigational links can also be used within a single document to assist users to jump to different parts of the document more easily.

On the other hand, hyperlinks can also be considered as recommendations to users. Links in one document can be used to direct users to other documents that address a similar topic of or provide additional information about the content of the former. A recommendation hyperlink conveys the message that two documents are related semantically. These links are therefore particularly useful in clustering Web documents based on their topics (Zhang et al., 2008). This type of link is commonly seen in the blogosphere nowadays, as bloggers refers to blog posts by other bloggers to express their opinions and ideas.

Hyperlinks allow individual Web documents to be connected to each other. The overall link structure of the Web is widely used in different applications such as crawling, ranking, retrieval and clustering of Web documents. Crawler programs of search engines rely mainly on the hyperlinks to retrieve documents one after another for indexing purpose. The PageRank algorithm, the algorithm behind the popular search engine Google, is probably the most well-known algorithm that exploit the link structure of the Web to assess a certain quality of the documents (Brin and Page, 1998). PageRank considers each link from one document to another as a 'vote' given by the former to the latter. The more votes a document receives from other documents, the higher quality the document should have. The HITS algorithm mentioned in the previous chapter is another example of analysing the link structure of the Web to rank documents.

The simple mechanism of linking one document to another, when deployed on the open and universal platform of the Web, gives rise to a very complex network structure. Broder et al. (2000) are the first to study the topology of the network of documents on the Web. They deduce from their data that the Web features a bow tie shape (see Figure 6.1, which mainly consists of a strongly connected component (SCC) at the centre of the Web, a set of documents guiding users into the SCC, and a set of documents guiding users out of the SCC. Other studies show that the in-degrees and out-degrees of the documents on the Web follows

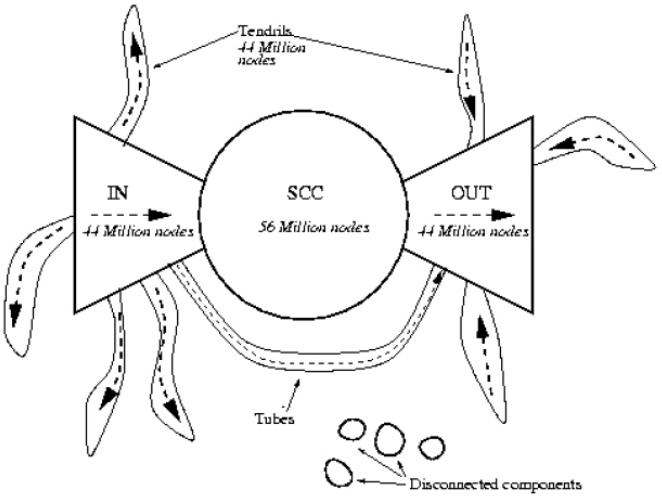


FIGURE 6.1: The shape of the Web graph as described by Broder et al. (2000).

the power law (Barabási et al., 2000; Kleinberg et al., 1999)—there are very few documents having high in-degrees or out-degrees but a lot of documents having low in-degrees or out-degrees, and that the diameter of the Web is surprisingly small given its huge scale (Albert et al., 1999), such that a user can go from one hypertext document to another with very few clicks on the Web relative to the enormous number of documents on the Web.

6.1.2 Implicit Links

While Web users are supposed to mainly rely on hyperlinks to navigate from page to page on the Web, following the hyperlinks is not the only way for the users to discover other relevant and useful documents. In the early days of the Web, online directories, such as the Yahoo! Directory and other 'hub pages' that contains a lot of hyperlinks to documents grouped under different categories, point users to documents that address similar topics but may not be connected to each other by hyperlinks. Nowadays, when directories are unable to keep up with the huge number of new documents appearing every day, users rely heavily on Web search engines such as Google and Yahoo! to look for documents that are relevant to their information needs. There are also personalised recommendation systems (Montaner et al., 2003) that suggest related documents to users based on their browsing behaviour, preferences and similarity between the interests of users. As a result, it is possible that there are a lot of Web documents that are not linked with each other but are often visited by users together because they contain information

about the same or similar topics.

There are some studies in the literature that investigate this kind of implicit links—not being connected by hyperlinks but are deemed related by users as manifested in their browsing behaviour—between Web documents. For example, Xue et al. (2003), while studying search within a single Web site (a small Web), propose to apply association rule mining to Web logs in order to find out pairs of pages that are frequently visited within the same session of browsing by users of the Web site. The authors define implicit links as links between two of this kind of pages. The authors find out that only about 11% of the discovered implicit links are existing hyperlinks. They believe that this result suggests 'the gap between the designer's expectation and the visitor's behaviour'. Experiments done by the authors show that running the PageRank algorithm on the link structure enhanced by implicit links gives better performance in retrieval.

Along a similar line of thought, Kazienko and Pilarczyk (2006) propose to use implicit links found in Web logs to evaluate the quality of existing hyperlinks within a Web site. Again using association rule mining, the authors discover two types of relations between documents: (1) $d_i \Longrightarrow d_j$: users visiting d_i are likely to visit d_j, and (2) $d_i \Longrightarrow\!\!\sim d_j$: users visiting d_i are not likely to visit d_j. They then use these discovered rules to judge whether existing hyperlinks are useful or not. In other words, the authors propose to use information about user browsing behaviour as an implicit feedback to assess the utility of existing hyperlinks.

In addition to Web logs, search engine query logs also offer similar user browsing information for the study of implicit links. Shen et al. (2006) present a method to generate implicit links from query logs for the purpose of Web page classification. They propose that two documents in a search result are linked by an implicit link if they are both chosen (clicked) under the same query submitted by the same user. The authors find that making use of these implicit links improves results of Web document classification, due to the fact that implicit links tend to connect documents of similar topics.

We believe that implicit links do not only help improve computational processing—such as ranking, classification and clustering—of Web documents. They should also be very useful in directly facilitating navigation of Web users. The fact that implicit links improve performances in retrieval and clustering suggests that they are good recommendation links that connects documents that are highly related to each other. By providing this kind of links in addition to explicit hyperlinks to the users when they are browsing, the users can then be provided with more

relevant information. Hyperlinks in a Web document are in general created by the author of the document. It is conceivable that these hyperlinks may not be adequate from the perspective of the readers of the document. For example, the author may not aware of some of the Web documents that are highly relevant to his/her own document, and thus fail to create such hyperlink. It may also because that some highly relevant documents are created by rivals of the author and they may be competing to attract more readers. In this case a hyperlink between these documents is not likely to exist, and an implicit link will prove to be very useful in such scenarios.

6.2 User-induced Links in Folksonomies

A folksonomy is actually quite similar to Web logs and search engine query logs in the sense that it also contains information about the preferences of users under different topics (represented by the tags contributed by the users). However, there are some differences between a folksonomy and a Web log or a search engine query log. Web logs and query logs do not necessarily show positive preferences of the users. In particular, we cannot be sure whether a user is interested in a document simply based on the fact that he/she has visited it or has clicked on it after submitting a query. A user may read a document only to find that it is not related or useful to what he/she has expected. On the contrary, in collaborative tagging users assign tags to documents usually because they are interested in it, want to share it or find it useful for later use. In other words, documents tagged by a user in a folksonomy can always be considered as interesting to the user.[3] It is thus more convincing to say that two documents are interested by a user when they both appear in a user's collection in a folksonomy than when they are both visited by the user or when they are both clicked on by the user after a search query. We believe folksonomies should be a better place for studying implicit links between documents.

Implicit links can be discovered in a folksonomy by different methods. For example, two documents that have both been assigned the tag `photography` by a large number of users can be considered as related to each other, and can be considered to be connected by an implicit link. In addition, given the large number of tags that have been assigned to the documents in a folksonomy, implicit links can also

[3]This is especially true when users cannot cast a 'negative' tag or vote to a document in most existing systems. Therefore the fact that a document appears in a user's collection is a strong evidence that he/she is interested in that document.

be found by calculating the similarity between the sets of tags assigned to the documents. For example, Markines et al. (2008) describe a system called GiveALink, which involves a global semantic similarity network to capture relationships among resources, and suggest that semantic similarity can be treated as an alternative way of navigating the Web by suggesting users to visit a page similar to the one being visited.

In other words, implicit links between documents in a folksonomy can be discovered by mainly two different approaches: (1) examining the tags that have been assigned to the documents, or (2) analysing the collective behaviour of the users who have tagged the documents. As implicit links in a folksonomy are resulted from the collaborative tagging activities of the participating users, we call them ***user-induced links***. In the following sections, we will discuss in detail these two approaches of discovering user-induced links.

6.2.1 Tag Similarity of Documents

The first approach of discovering user-induced links in a folksonomy is to calculate the pair-wise similarity of the documents and single out pairs of documents that achieve a certain level of similarity. The similarity between two documents can be measured using many different approaches. Given that documents are characterised by words, similarity is most naturally determined by comparing the set of keywords that are deemed representative of the content of the documents. Such a set of keywords can be extracted by stop-words filtering and weighting schemes such as TF-IDF (Salton and Buckley, 1987). A straightforward method of measuring similarity is to use the Jaccard coefficient:

$$Sim(T_a, T_b) = \frac{|T_a \cap T_b|}{|T_a \cup T_b|} \tag{6.1}$$

where T_a and T_b are the sets of keywords of documents a and b respectively.

However, such simple measure does not take into account the importance of different keywords. It is natural that certain keywords are central to the content of a document such that these keywords should be given more considerations. In the information retrieval literature, documents are usually characterised by term vectors (Manning et al., 2008) in a vector space. A term vector is a vector whose elements indicate the importance of the chosen keywords to the document. Similarity between two documents can be measured by using the cosine similarity

calculated on the two respective term vectors (reproducing Equation 4.8):

$$csim(\boldsymbol{v}_1, \boldsymbol{v}_2) = \frac{\boldsymbol{v}_1 \cdot \boldsymbol{v}_2}{||\boldsymbol{v}_1|| \times ||\boldsymbol{v}_2||}$$

and $\boldsymbol{v}_1, \boldsymbol{v}_2 \in \mathbf{R}^n$.

Alternatively, one can also consider a document as characterised by a tuple W_d, which involves a set T_d of tags and a weighting function w that maps a tag to its normalised weight representing its importance to the document:

$$W_d = (T_d, w_d) \tag{6.2}$$

where

$$T_d = \{t|\exists u, (u, d, t) \in R\} \tag{6.3}$$

$$w_d(t) = \frac{|\{(u, d, t)|\exists u \text{ s.t. } (u, d, t) \in R\}|}{|\{(u, d, t')|\exists u \text{ s.t. } (u, d, t') \in R\}|} \tag{6.4}$$

By using this representation of a document, we introduce two different similarity measures for assessing the similarity of two documents. The first similarity measure is a weighted version of the Dice coefficient (Ker and Chang, 1997) that is widely used in set comparison:

$$Sim_w(T_a, T_b) = \frac{\sum_{t \in T_a \cap T_b} w_a(t) + w_b(t)}{\sum_{t \in T_a} w_a(t) + \sum_{t \in T_b} w_b(t)} \tag{6.5}$$

which can be simplified to

$$Sim_w(T_a, T_b) = \frac{\sum_{t \in T_a \cap T_b} w_a(t) + w_b(t)}{2} \tag{6.6}$$

since $\sum_{t \in T_a} w_d(t) = \sum_{t \in T_b} w_d(t) = 1$ because weights of the tags are normalised. This weighted Dice coefficient returns a higher similarity value if the two documents share keywords of higher importance (larger weights).

The second similarity function we introduce here is based on the normalised discounted cumulative gain (NDCG) (Järvelin and Kekäläinen, 2002). NDCG is a performance measure mainly used in information retrieval research to evaluate rankings of documents according to their relevance. It measures how good a ranking algorithm is in assigning suitable ranking to relevant documents. For example, if we have three documents $\{d_1, d_2, d_3\}$ whose relevance scores are $(3, 2, 1)$ respectively (higher score means more relevant), then a ranking (d_1, d_2, d_3) will attain

a higher NDCG than another ranking (d_3, d_1, d_2), because the first one assigns higher ranks to documents that are more relevant.

Here, we borrow the idea of NDCG to measure the similarity of two documents based on their tags and the associated weights. Assume that we have two documents d_1 and d_2, and we now want to measure how similar d_2 is to d_1. For d_1, we have a list of tags organised in descending order of their weights, $(t_1, t_2, ..., t_n)$, whose weights are $(w_{d_1}(t_1), w_{d_1}(t_2), ..., w_{d_1}(t_n))$. We treat the tags of d_1 as items to be retrieved and ranked, and treat their weights as their relevance scores. As a result, the list of tags of d_2 can be considered as a ranking result produced by some ranking algorithm to reproduce the list of tags of d_1 as accurately as possible. In this way, two documents with the same set of tags and same ordering according to their weights will achieve an NDCG of 1, two documents which share no tags at all will result in an NDCG of 0. It should be noted that unlike the weighted tag similarity the NDCG similarity measure is asymmetric.

Formally, calculating the NDCG similarity of d_2 to d_1 requires several steps. Firstly, lists of tags of the two documents are prepared, with the tags ordered in descending order of their weights:

$$l_{d_1} = (t_{d_1,1}, t_{d_1,2}, ..., t_{d_1,n}) \tag{6.7}$$

$$l_{d_2} = (t_{d_2,1}, t_{d_2,2}, ..., t_{d_2,n}) \tag{6.8}$$

Secondly, the discounted cumulative gain (DCG) at position p is calculated by:

$$DCG_p = w_{d_1}(t_{d_2,1}) + \sum_{i=2}^{p} \frac{w_{d_1}(t_{d_2,i})}{\log_2 i} \tag{6.9}$$

Thirdly, we need the ideal discounted cumulative gain (iDCG) at position p, which is the DCG at position p when tags are ranked exactly according to their weights (the ideal case). It is used to normalise the DCG obtained using the above equation such that the final NDCG value varies in the range of 0 to 1.

$$iDCG_p = w_{d_1}(t_{d_1,1}) + \sum_{i=2}^{p} \frac{w_{d_1}(t_{d_1,i})}{\log_2 i} \tag{6.10}$$

Finally, the NDCG value is calculated by simply obtaining the ratio between DCG and iDCG:

$$NDCG_p = \frac{DCG_p}{iDCG_p} \tag{6.11}$$

Given these similarity measures, it becomes possible to discover user-induced links between pairs of documents that are similar to each other. One important issue in using similarity to discover implicit links is that we are likely to discover a huge number of implicit links. This is because it is common for documents to share one or two very general tags even though they may not be particularly related to each other in terms of their content. These documents will nevertheless achieve non-zero similarity. Hence, a threshold value of similarity should be specified in order to narrow down the results to a reasonable and useful set of implicit links. In summary, the process of discovering user-induced links using one of the similarity measures can be represented by a function that takes the set of documents, the chosen similarity measure and the similarity threshold as parameters:

$$G_s(D, Sim, threshold) = \{(d_i, d_j)\} \tag{6.12}$$

6.2.2 User Preferences

The second approach of discovering implicit links involves finding out pairs of Web documents that have both been tagged by the same group of users, probably with the restriction of under the same tag or same set of tags. The method for identifying such pairs of Web documents can in fact be readily borrowed from the data mining research area. The task of mining association rules from large databases Agrawal et al. (1993) aims at identifying implicit patterns within a large database of transactions. In traditional association rule mining, a classic example would be that people who buy bread and butter in the supermarket are very likely to buy milk as well. Borrowing such idea to the context of collaborative tagging, the problem becomes one of identifying pairs of Web documents such that when users have tagged one of them they are very likely to tag the other one as well. In other words, we can use the technique of association rule mining to discover these user-induced links.

Note that when considering user preferences, we are ignoring the content of the documents as well as the tags assigned by users to the documents. We are only judging the relations between documents based on the preferences of the users. This approach is actually very similar to the user-based document networks we mentioned in Chapter 4, in which documents are characterised by the users who have assigned a particular tag to them, and the similarity between their sets of users determine how closely related two documents are. As for discovering user-induced links here, instead of merely computing the similarity of their sets of users,

we use association rule mining such that the direction of a link, i.e. where a link originates and where it ends, can also be determined.

Formally, let $D = \{d_1, d_2, ..., d_n\}$ be a set of Web documents, and C be a database of document collections. Each $c_u \in C$ represents the set of documents that have been tagged by the user u. In traditional association rule mining, let X and I_k denote sets of items, rules can assume the form of $X \Longrightarrow I_k$, meaning that the presence of X in a certain transaction implies a high probability of the presence of I_k in the transaction. However, in the case of identifying user-induced hyperlinks, it is not very helpful to discover something like 'd_1, d_2 and d_3 should altogether have a link to d_4', as links should be originated from a single document to another single document. Hence, we will focus on discovering association rules in the form of $d_i \Longrightarrow d_j$.

Two major concepts in association rule mining are *support* and *confidence*. In our context, support of a set of documents is defined as the proportion of collections in the database that contain the set of documents:

$$supp(X) = \frac{|\{c|\forall d \in X, d \in c\}|}{|C|} \tag{6.13}$$

In general, we aim at discovering rules that have large supports. This is because a larger support implies that the rule involves documents that are more popular among the users. Therefore rules of larger supports will find themselves more useful in the future.

Confidence of a rule $d_i \Longrightarrow d_j$, on the other hand, is defined as the proportion of collections in the database in which the rule is correct:

$$conf(d_i \Longrightarrow d_j) = \frac{supp(\{d_i, d_j\})}{supp(\{d_i\})} \tag{6.14}$$

In general, we also want the confidence of a rule to be as high as possible. The confidence of a rule actually corresponds to the extent to which the rule is a valid one. A rule that has a higher confidence would mean that it would be more likely to obtain a correct result when the rule is applied. In the context of discovering user-induced links in folksonomies, a higher confidence means that the user-induced link is deemed appropriate by more users and therefore it is more likely that such a link would benefit other users as well.

Note that similar to the case of NDCG similarity, user-induced links discovered by using association rule mining are not symmetric. The existence of the rule

$d_i \Longrightarrow d_j$ does not imply the existence of the rule $d_j \Longrightarrow d_i$ because the two rules would have different levels of support and confidence. In summary, the process of discovering user-induced links in a folksonomy using association rule mining can be represented by the following function:

$$G_u(D, C, min\_supp, min\_conf) = \{(d_i, d_j)\} \tag{6.15}$$

6.3 Analysis of User-induced Links

By using the two methods described above, we identify user-induced links in data collected from Delicious and compare them with existing hyperlinks in terms of several different aspects. In performing the analysis and comparison, we focus on whether the links (including existing hyperlinks and user-induced links) can be considered as good recommendation links. While it can be a subjective judgement of whether a link makes good recommendation to a user, we believe there are several aspects of a link that we can study and measure to answer the question. These aspects include whether a link connects two documents from the same domain/Web site, the similarity between documents on the two ends of a link, and whether users are equally interested in the linked documents. We will perform our analysis along these dimensions.

6.3.1 Data Preparation

To conduct the experiments, we again rely on the data sets collected from Delicious as described in Chapter 3 (data sets involving over 1,000,000 unique users, over 100,000 unique documents, and over 800,000 unique tags collected based on 135 seed tags). The data sets collected by crawling Delicious based on a set of seed tags are very suitable for our experiments here. This is because Delicious contains a huge amount of data, and we are not likely to obtain many user-induced links if we conduct our study on a randomly collected data sets due to data sparsity. By focusing at each time on a set of documents all assigned a particular tag, we reduce the diversity of the tags found in the documents as well as the diversity of users who have tagged the documents.

In addition to the users and tags associated with the documents, we also need the existing hyperlinks embedded in the documents for comparison. To obtain the existing link structure among the collected documents, we download each of

Association Rule		**Weighted Similarity**		**NDCG Similarity**	
Confidence	Links	Threshold	Links	Threshold	Links
> 0.10	75	> 0.50	11,724	> 0.50	15,294
> 0.15	39	> 0.55	7,371	> 0.55	10,897
> 0.20	20	> 0.60	4,279	> 0.60	9,969
> 0.25	13	> 0.65	2,696	> 0.65	8,981
> 0.30	10	> 0.70	1,545	> 0.70	8,050
> 0.35	10	> 0.75	667	> 0.75	7,356
> 0.40	9	> 0.80	516	> 0.80	6,806
> 0.45	7	> 0.85	408	> 0.85	6,466
> 0.50	6	> 0.90	366	> 0.90	6,240
		> 0.95	355	> 0.95	6,125

TABLE 6.2: Average number of induced links generated for each tag data set by different methods using different parameters. The figures are averaged over the 130 data sets collected from Delicious.

them and parse the HTML source code to identify their outgoing links. Since our experiments focus on the characteristics of the documents on the two ends of a link (both the source and the destination), we do not consider links that points to documents not in our data. In addition, it is possible that there are more links between the documents than we have collected. This is because hyperlinks can also be embedded in more sophisticated ways in a document by using for example Javascript or Flash content. At the end of this process we have 56,900 links. The maximum number of outgoing links for a document is 58, and the maximum number of incoming links is 240.

6.3.2 Results

We identify user-induced links between the documents in each of the 135 data sets by the two proposed methods using different parameters. For the similarity approach, we vary the similarity threshold. As this approach tends to return a lot of user-induced links, we only focus on links between documents with similarity of at least 0.5. For the association rule mining approach, we set the minimum support at 100 and vary the minimum confidence level. We find that very few user-induced links achieve a confidence level of 0.5 or above.

Table 6.2 shows the number of user-induced links generated by using different methods and parameters. An obvious difference between the different methods is that the use of tag similarity generates far more user-induced links than the use of association rule mining. This actually reflects a major difference between the two methods. In using similarity, we compare the tags of different documents, since

we focus on a group of documents with a particular tag at a time, the documents are already confined to a single (though very general) topic. As a result, the diversity of tags found in this group of documents is far smaller than the diversity of users who are interested in these documents, thus resulting in a much higher 'tag similarity' than 'user similarity' among the documents.

While we note that weighted similarity and NDCG similarity have a high correlation ($r \approx 0.85$), there are actually some differences between the two. Although both similarity measures consider the weights of the tags, NDCG puts much more emphasis on matching tags that are most important. Consequently, if two documents have their a few most popular tags in the same order, they are very likely to attain a higher value in NDCG similarity than in weighted similarity. As a result, we see that NDCG similarity gives us a lot more induced links than weighted similarity.

6.3.2.1 Number of Same-Domain Links

One important function of hyperlinks is to allow users to navigate from one hypertext document to another, especially those within the same Web site. Arguably, it would be more beneficial to a user if links point to some documents external to the current Web site, which should provide relevant information different from that available in the current one. For example, links from a blog post in one blog to blog posts in another blog would be more informative in general than links to blog posts within the same blog. Hence, it would be interesting to compare this aspect in existing hyperlinks between the documents and the links induced from the tagging behaviour of Web users.

For each of the existing hyperlinks and the induced links, we check whether the documents at the two ends of the link are from the same domain. We do this by comparing their URLs and see if they have the same domain name. For example, a test on `http://developer.apple.com/` and `http://support.apple.com/` will be positive as they are both under the domain name of `apple.com`. We note that, however, this may overestimate the number of links connecting documents from the same Web site. This is because two URLs having the same domain name but different sub-domain names may well be referring to two different Web sites. For example, we may want to consider a blog at `http://userA.blogspot.com/` and another blog at `http://userB.blogspot.com/` as two different Web sites, although they are both under the same domain. In practice, these subtle differences

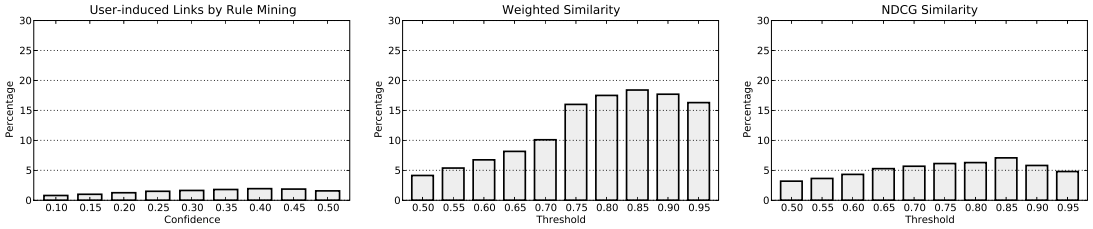


FIGURE 6.2: Percentage of user-induced links connecting documents from the same domain.

may be difficult to distinguish from one another when automatic processing of the URLs is involved. Nevertheless, since we compare the different types of links on the same basis, this should not be considered as a problem.

Figure 6.2 shows the percentage of links that connect documents from the same domain for user-induced links generated by using the three different methods. We note that for existing hyperlinks about 33% of them are between documents from the same domain, and the probability of having such a same-domain link in our data sets is about 15%. Firstly, we see that only about 1-4% of user-induced links generated by association rule mining are connecting documents from the same domain. This is much lower than that of existing hyperlinks and by chance, suggesting that users are very unlikely to be interested in multiple documents from the same domain.

The graphs of the links generated by similarity measure seem to guess that there is a difference between weighted similarity and NDCG similarity. However, taking the different number of links generated in the two cases into consideration, this difference is only due to the different distribution of links among the similarity level. The number of links generated by NDCG similarity that attain a similarity level of 0.95 is greater than that generated by weighted similarity that attain a similarity level of 0.60. This shows that NDCG is less fine-grained than weighted similarity, and it is relatively easier to achieve high similarity in NDCG. The graph for weighted similarity suggests that links in which documents are more similar are more likely to be from the same domain. When we pick some of these links for further investigation, we see that many of these links are between a series of documents addressing the same topic in a blog, or tutorials of highly related applications. Nevertheless, compared to existing hyperlinks, there are much fewer user-induced links that connect documents from the same domain.

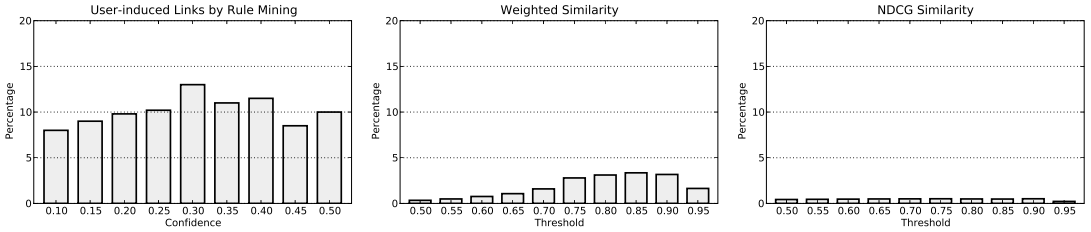


FIGURE 6.3: Percentage of user-induced links that are existing hyperlinks.

6.3.2.2 Coincidence between Different Link Types

In addition to examining the domains of linked documents, another way to study the usefulness of user-induced links is to see whether such links already exist between the documents. If user-induced links coincide with existing hyperlinks, it would suggest that users are satisfied with the existing hyperlinks and do not pay much attention to documents that are not linked. On the other hand, if user-induced links are mostly new, it means that there are user interests and perspectives that existing hyperlinks have not captured.

Figure 6.3 shows the percentage of user-induced links that are the same as the existing hyperlinks. The graphs seem to show that user-induced hyperlinks generated by association rule mining are more likely to be existing hyperlinks, and that those generated by NDCG similarity are less likely to be so. However, we again have to take into account the different numbers of user-induced links generated in different cases. Since there are a lot more user-induced links based on similarity than those based on association rule mining, it is understandable that the former coincide much fewer existing hyperlinks.

Path Length	**Frequency**	**Percentage**
∞	6,442	89.26%
1	439	6.08%
2	132	1.83%
3	57	0.79%
4	42	0.58%
5	29	0.40%
> 5	76	1.05%

TABLE 6.3: User-induced links and the lengths of the shortest paths between the documents concerned.

The result for induced links generated by association rule mining is particularly interesting. This is because given the relatively few user-induced links in this case, the overlap between these and existing hyperlinks is at most about 13%. This

shows that a hyperlink does not necessarily connect documents both of which users find interesting or useful. In other words, users tend to find out related documents by other means because there are no hyperlinks between them. It is possible that two documents are not directly linked but can be reached by two or more hops on the Web graph. However, as shown in Table 6.3, only a very tiny portion of documents that are not directly linked can be reached by more hops.[4] In addition, one may suggest that users do not tag both documents connected by a link simply because of the existence of the link: it is sufficient to save one of them which will lead the user to the other. However, given that all these documents have been tagged by some users, it suggests that all these documents deserve to be bookmarked for future retrieval.

To get a better understanding of the user-induced links, we look into documents that are connected by these links but not by existing hyperlinks. We find that many of the user-induced links are (1) between blog posts of highly related topics, (2) news articles on the same topics, (3) Web sites offering applications of similar functionalities, and (4) Q&A pages of some portal sites. In all these cases, there are some reasons that hyperlinks do not exist. For example, the author of a document may not be aware of other related documents (as in 1 and 2), or two Web sites are competing for readership because they offer similar content (as in 3), or the system is not designed to be aware of the similarity of its content (as in 4).[5]

The results of similarity-based user-induced links are less surprising given the very large number of links generated. However, they do show that existing hyperlinks are very inadequate when they come to recommend related documents to the users. There are just much more related documents out there than those to which hyperlinks within a document point to. Of course, it would not be practical for a document to be linked to all of, for example, the 10,000 documents that contain related materials. Nevertheless, the results suggest that there are clearly room for improvement for existing hyperlinks.

6.3.2.3 Similarity and User Preferences

The two approaches for generated user-induced links are completely different from each other. Association rule mining concerns with the preferences of the users,

[4]It is possible that a path exists by traversing documents that are not within our data sets. However this is beyond our scope as that requires the knowledge of the global Web graph.

[5]Case 4 is likely to be found on FAQ documents provided by authors. In many user-contributed Q&A sites like Yahoo! Answers, similar questions and answers are usually recommended to the users by the systems.

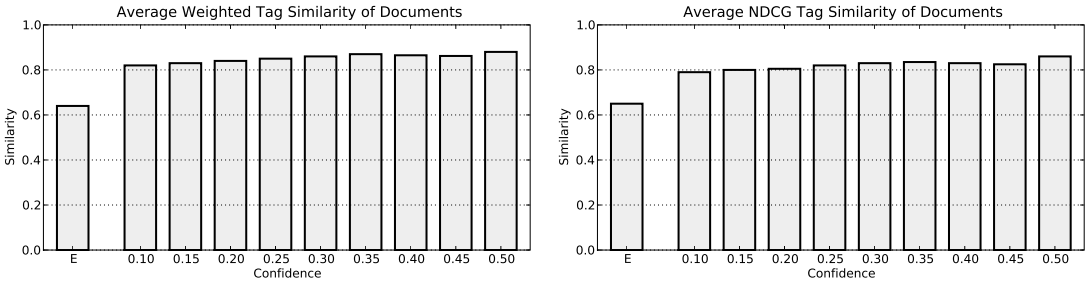


FIGURE 6.4: Average similarity of pairs of documents on the two ends of user-induced links generated by association rule mining. 'E' refers to the results obtained for existing hyperlinks.

and a link is generated if enough users are interested in two documents, regardless of the similarity between them. On the other hand, the similarity-based approach generates links based on the tags assigned to the documents, regardless of whether there are many users interested in the documents on the two ends of the links. In this section, we investigate whether the links generated by one method satisfy the requirement of the other method.

Figure 6.4 graphs the similarity between documents connected by user-induced links generated by association rule mining. We can see that the pairs of documents all attain high similarity in the two similarity measures. It shows that pairs of documents that are interested by many users are actually very similar to each other with respect to the tags assigned to them, which are indicative of their topics. We also calculate the similarity between documents connected by existing hyperlinks for reference. As shown in Fig. 6.4, existing hyperlinks achieve about 0.62 in both similarity measures, which is much lower than those achieved by user-induced links. This result for existing hyperlinks is expectable because many of them serve navigational purposes and therefore it is not uncommon for their sources and destinations to involve content of different topics (e.g. a link back to the front page of a Web site or a link from a blog post to the profile of the blogger).

Next, we investigate whether similarity-based user-induced links connect pairs of documents that are interested by many users. Figure 6.5 graphs the number of users that have tagged both documents on the two ends of a link. For both existing hyperlinks and similarity-generated user-induced links, we see a power law-like distribution of the number of overlap users. In other words, there are only a very small number of links that connect documents both of which are interested by a large number of users, and there are a large number of links for the opposite case. Hence, contrary to the findings in user-induced links generated by association rule

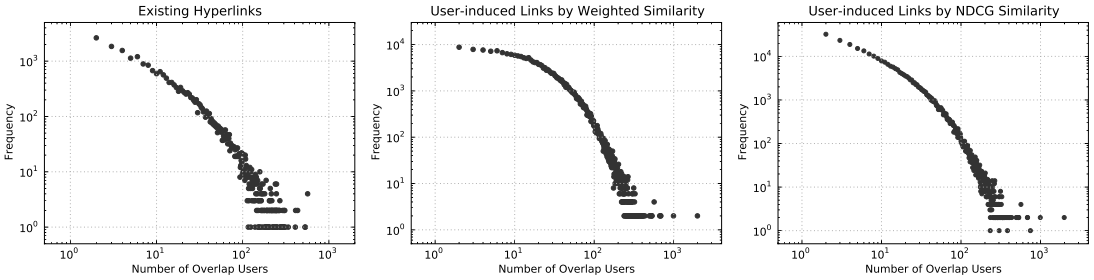


FIGURE 6.5: Number of users that have tagged both documents on the two ends of a link. It can be seen that relatively very few users express explicit interests in documents linked by existing hyperlinks or user-induced links generated by tag similarity.

mining, high similarity does not guarantee high user preferences. It may suggest that users are much more selective and do not only consider the similarity between two documents. It may also suggest that given the large amount of information users choose to focus on a small number of documents that are related and are useful from their own perspectives.

Putting the findings described in this section together, we can see that explicit user preferences (association rule mining of user collections) represent better filters of useful relations between documents that similarity measures. The former satisfies both user preferences as well as topical similarity between the documents, whereas the latter does not necessarily produce links that are interested by many users. Nevertheless, both approaches can be considered as useful means for identifying implicit relations between documents that are not captured by existing hyperlinks, as our experiments show that user-induced links offer much new information that cannot be found in existing hyperlinks.

6.4 Tag Prediction

The analysis of user-induced links shows that links generated by association rule mining of user collections usually connect documents that are highly related to each other as judged by the similarity between their tags. This result inspires us to use user-induced links to predict tags of a document. Given that documents connected by user-induced links have highly similar sets of tags, aggregating the tags of documents linking to a chosen document is probably be a good way of predicting the tags of this document. For example, if all documents linking to this

document have been assigned the tag `photography`, it becomes very probable that this document can also be suitably described by this tag.

Tag prediction has not received attention until only very recently. However, tag prediction can be very useful in many scenarios. Heymann et al. (2008b) remark that tag prediction can enhance a collaborative tagging system in several different ways. In particular, by using tag predicting to enrich the set of tags of documents, we can increase recall of single tag queries. This is valuable when many documents in a folksonomy have only been assigned tags by one or two users (a characteristic of power law distribution). Of course, even when there are many users who have assigned tags to a document, the tags can also be very general and not specific enough for retrieval.

There are a few proposals in the literature for tag predictions. Heymann et al. (2008b) experiment with identifying occurrence relations between tags by using association rule mining with the objective of expanding the set of tags of a document. For example, by applying association rule mining algorithms on Delicious data the authors discover rules such as `debian` $\Longrightarrow$ `linux`, meaning that a document tagged by `debian` is very likely to be tagged by `linux` as well. Hence, if a document has been assigned `debian`, the tag `linux` can also be attached to it, so that other users searching with `linux` as a query tag can also retrieval this document. On the other hand, Budura et al. (2009) present a more sophisticated method of tag prediction that takes into account the neighbours of a document in a graph, such as a graph of hyperlinked documents or a graph of citation between papers. The method is based on a scoring method that propagates the weights of different tags within the graph of documents under consideration.

We note several differences between the tag prediction method we propose here and methods in the literature including those mentioned in the above paragraph. Firstly, previous studies on tag predictions mostly focus on expanding an existing small set of tags for a document, instead of predicting its tags when it has not yet been assigned any tags. Secondly, our method considers user-induced links instead of only existing hyperlinks or citation links to predict tags of a document. While the methods described in the literature are successful in expanding the existing set of tags of the documents, their performance can be affected by the initially assigned tags of a document or by the quality of existing hyperlinks. Methods that aim at expanding a set of tags work well when the initial tags are specific ones. If the first few tags assigned by the users are very general, these tag prediction methods will be less accurate as there are much more possibilities. For example, predicting the

tag `debian` would be relatively easy given the presence of the tag `linux`. However, it is not that straightforward the other way round, as there are clearly many other different kinds of Linux systems. In addition, as we have shown earlier in this chapter, hyperlinks do not always connect documents of similar topics. When there are more navigational links than recommendation links in the Web graph, performance of tag prediction based on existing hyperlinks will then be affected.

We hypothesise that user-induced link are more useful in predicting tags of a document. In addition, this approach would not require an initial set of tags for the document to start with. In the following sections, we first describe our proposed method, and then present and discuss our experiments in which we compare the usefulness of existing hyperlinks and user-induced links. Note that we will focus only on user-induced links generated by association rule mining of user preferences. This is because there would be no point in predicting tags based on a document's neighbours when all the neighbours are linked to the document because they have similar tags.

6.4.1 Proposed Method

To predict the tags of a certain document, we first need to identify the other documents that have a link to this document. Let $G = (D_G, L_G)$ be a graph with a set D_G of vertices representing documents and a set L_G of arcs representing links between the documents. We consider both a graph G_w of existing hyperlinks and a graph G_u of user-induced links. For a document d_x in the graph, the set of documents that have a link to d_x is given by:

$$P_G(d_x) = \{d | (d, d_x) \in L_G\} \tag{6.16}$$

Our hypothesis is that documents in $P_G(d_x)$ contain information related to the content of d_x, and therefore the tags of the documents in $P_G(d_x)$ should also be applicable to d_x. We can aggregate the tags of these documents and use them to predict the tags of d_x. We consider two different methods of aggregating the tags of documents in $P_G(d_x)$. Firstly, we consider a simple averaging method: we come up with a set of tags with their weights equal to the average of their weights in documents in $P_G(d_x)$. Let $W^a_{d_x} = (T^a_{d_x}, w^a_{d_x})$ represents the prediction (superscript a means average aggregation), where $T^s_{d_x}$ is the set of tags and $w^s_{d_x}$ is a function

that returns the weight of the tags. Our first method of aggregation is given by:

$$T^a_{d_x} = \bigcup_{d \in P_G(d_x)} T_d \tag{6.17}$$

$$w^a_{d_x}(t) = \frac{1}{|P_G(d_x)|} \sum_{d \in P_G(d_x)} w_d(t) \tag{6.18}$$

In addition, by assuming that an induced link of higher confidence will connect a more related document to d_x, we also consider a slightly sophisticated method of aggregation by taking the confidence of the link into account. Let $conf(d_1 \Longrightarrow d_2)$ be the confidence of the user-induced link from d_1 to d_2. Our second method of aggregation is given by $W^w_{d_x} = (T^w_{d_x}, w^w_{d_x})$ (superscript w means weighted aggregation), where

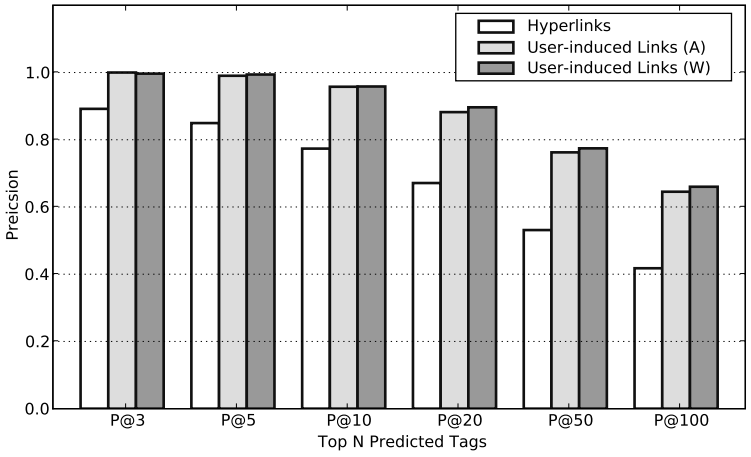
$$T^w_{d_x} = \bigcup_{d \in P_G(d_x)} T_d \tag{6.19}$$

$$w^w_{d_x}(t) = \frac{\sum_{d \in P_G(d_x)} w_d(t) \times conf(d \Longrightarrow d_x)}{\sum_{d \in P_G(d_x)} conf(d \Longrightarrow d_x)} \tag{6.20}$$

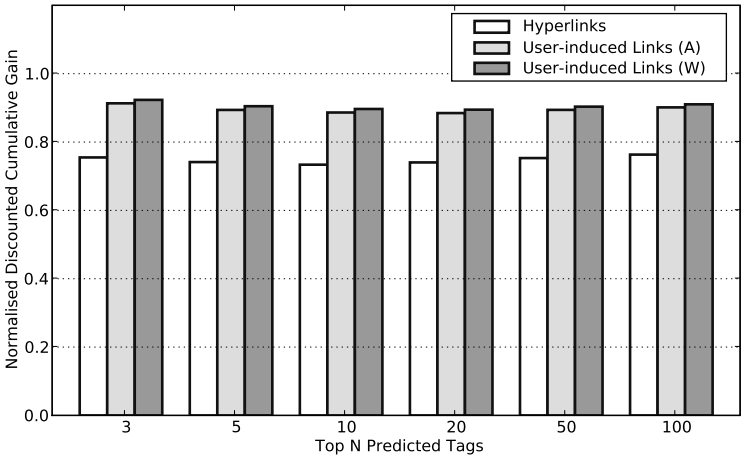
Note that our proposed method of predicting the tags of a document is similar to the k-nearest-neighbour algorithm for classification, in which the class label of an item is determined by those that are closest to it, except that in our case the number of neighbours of a document is not fixed and depends on the number of user-induced links that have this document as destination. In other words, tag prediction can also be considered as a classification problem. Performance of user-induced links in tag prediction is therefore indicative of their usefulness in Web document classification.

6.4.2 Experiment

From our data sets, we select documents that have at least 5 incoming user-induced links and at least 5 incoming hyperlinks from other documents to ensure that we have enough data for the prediction process. After the filtering process we obtain a total of 1,241 documents satisfying the above conditions. On average a document in the set has 9 incoming user-induced links and 14 incoming hyperlinks. We use the average aggregation method to generate predictions from hyperlinks (as they do not have any confidence values), and use both average and weighted aggregation method to generate predictions from user-induced hyperlinks.



(a)



(b)

FIGURE 6.6: (a) Precision levels of predictions at different number of tags. (b) Normalised discounted cumulative gain (NDCG) of the predictions. (A) means average aggregation of tags, and (W) means weighted aggregation of tags.

We measure the performance of the predictions by using NDCG as well as precision at the nth item. Precision at the nth item is calculated by measuring the precision of the first n tags, i.e. the top n tags with largest weights, in the prediction. On the other hand, NDCG as a performance measure works effectively in the same way as described in Section 6.2.1. We use NDCG mainly to investigate whether the predictions are accurate in terms of the ordering of the tags. In our experiments, we use the tags assigned to the documents by the users in Delicious as the ground truth.

Figure 6.6(a) shows the precision levels of the predictions for different values of n. We can see that predictions based on user-induced links are significantly more accurate that those based on existing hyperlinks, with precision of 90% or higher

for the first 20 tags. The performance of using weighted aggregation gives slightly better results than using average aggregation. Note that the number of predicted tags is always larger than the actual number of unique tags assigned to the document in Delicious since we do not impose any threshold on the weight of the tags. Given the fact that precision decreases as we consider more and more tags in the prediction, it can be concluded that correct tags are usually given higher weights in the prediction than wrong tags. This is confirmed by the results given by the NDCG measure.

Figure 6.6(b) shows the NDCG values of the predictions when different number of top predicted tags are considered. Again, we see that predictions based on user-induced links attain significantly higher values than those based on existing hyperlinks, and that weighted aggregation gives slightly better results than average aggregation. Judging from the fact that the NDCG values of the predictions are always higher than 0.9, the user-induced links represent a good basis for predicting even the relative importance of different tags to a document. An interesting result is that the values of NDCG do not change much at different positions. They are more or less constant even we consider more tags in the predictions. This is in fact related to the popularity of the tags. We observe that the number of times the tags are used on a document usually follows the power law, with a few tags very popular among the users and a large number of tags that are only favoured by a small number of users. Hence, once the first few tags are correctly predicted a high NDCG value will be obtained, and subsequent correct or incorrect predictions will not change the value significantly.

6.5 Discussion

The study of user-induced links in this chapter reveals that implicit relations between Web documents can be discovered by examining user preferences and document similarity embedded in a folksonomy. We also show that user-induced links are very different from existing hyperlinks in several different aspects, including the proportion of links between documents from the same domain, the number of users interested in the documents and the similarity between the documents.

An important aspect of the Web revealed by this study is that, at least within a collaborative tagging environment, there is a big difference between the perspective of Web authors and that of Web readers (Bao et al., 2007). This can also be framed as a difference between the expectation of Web designers and the behaviour of

Web surfers, or even a difference between Web 1.0 and Web 2.0. Hyperlinks are supposed to provide users with recommendations of related documents, but it turns out that users find out interesting documents very often without the help of hyperlinks. This suggests that it is very desirable to complement the existing link structure on the Web with information of user preferences, which indicate between which documents hyperlinks should be added.

It is true that the user-induced links we discussed in this chapter are document-level links, as opposed to hyperlinks which are object-level links. In other words, a user-induced link only tells us whether a document should be linked to another document, instead of in which part of a document the link should be embedded. It would be interesting to investigate how we can identify the correct part of a document where we can insert a user-induced link. However, document-level links are still very useful as recommendations to the readers of a document, just as the list of 'further readings' provided at the end of a book, which is complementary to the cited references within the text. White et al. (2007) propose a system in which documents that are frequently visited by users submitting same or similar queries are shown along side the search results presented to a user. The authors find that such a system leads to 'more successful and efficient searching compared to query suggestion and unaided Web search'. The frequently visited documents in this system are actually equivalent to the destinations of the user-induced links studied in this chapter. In other words, there are evidence that even document-level links can be beneficial to Web users.

In fact, many Social Web applications already generate document-level implicit links among their own documents to facilitate their users. For example, when one views a video clip on YouTube, a list of related video clips are presented to the user in a side bar. While the exact mechanism by which these links are generated is unknown, it is speculated that it is generated by considering the documents' tags, titles, popularity, comments, etc. There is no doubt that these links improve user experience while they are visiting the site. However, these links are all confined to the content within a particular Web site. In some sense, these links are generated to attract the attention of the users such that the duration of visits can be kept as long as possible. Users are not presented with related information outside of the site. Hence, the methods described in this chapter is particularly useful because they tend to discover links between documents from different domains.

Nevertheless, we note that providing suggestions of hyperlinks to authors by using information of user preferences so as to allow them to improve their documents

may only be a limited solution to the problem. In the course of our study, as we have mentioned in Section 6.3.2.2, we discover that there are many user-induced links that are between Web sites that can be considered as rivals or competing for readership. Hence, it is not realistic to expect that authors of these Web sites would create such hyperlinks, even though if these links are suggested to the authors. In addition, it may as well be the authors' intention to limit the number of hyperlinks due to various reasons.

In the end, it may be worthwhile to consider something like an open hypermedia structure (Fountain et al., 1992) backed by a collaborative tagging system. In an open hypermedia system, links are treated as objects and they have their own identities and characteristics. In addition, links are not separated from the documents and are stored in independent databases usually called linkbases. By adopting the open hypermedia approach, links between different Web documents can be induced from the collective behaviour of the users, and can be maintained externally with respect to the documents involved. These links represent the perspective of the users on how documents on the Web should be linked to each other. There are also possibilities of working towards semantic links since documents have been assigned tags by users. For example, when we generate user-induced links between documents that are all assigned the tag `cooking`, induced links between these URLs can be described by the tag, giving the users an idea of why these URLs are linked. In other words, user-induced links have great potentials to be further studied.

6.6 Chapter Summary

In this chapter, we have studied how we can identify user-induced links, a form of implicit relations, between documents in a folksonomy by using two different approaches, namely tag similarity between documents and association rule mining of user preferences. We have showed that both user preferences and tag similarity can be used to generate many user-induced links. In particular, the use of association rule mining of user preferences generates very high quality user-induced links because they are both highly preferred by the users and connect documents that contain highly related content. This supports the hypothesis we mentioned at the beginning of this chapter that these implicit relations represent better recommendation links for users browsing the Web. We have also showed that user-induced links can be used to predict tags of documents with a very high accuracy, suggest-

ing that they can be utilised in document classification tasks as well. In summary, our study has revealed the differences between the perspectives of authors and that of readers on the Web. It also suggests that the collective user behaviour observed in collaborative tagging systems can be exploited to enhance the link structure on the Web.

With this chapter, we have come to the end of our analysis of the collective semantics of the three different types of entities in collaborative tagging, namely tags, users and documents. In the next chapter, we will give our conclusions of this thesis based on the analysis and findings described in this and the previous chapters. We will also discuss the implications and significance of our studies, and outline possible future research directions.

Chapter 7

Conclusions and Future Work

7.1 Conclusions

In this thesis, we have presented the notion of collective semantics in the context of the Social Web, which comprises a set of social networking applications that promote social interactions such as sharing of online resources on the Web. We focused on an exemplar of collaborative tagging systems, the social bookmarking site Delicious, as a case study of the Social Web. Collaborative tagging systems exhibit the most general form of social interactions on the Social Web, and involve all three important types of entity—tags, users and documents, such systems are therefore suitable subjects for our study. The main objective of this thesis was to study how collective user behaviours observed in a collaborative tagging system generate meaningful associations between the entities involved. It answered the question of how we can analyse these associations and in general the resultant network structures to uncover the semantics of these entities.

This thesis started with a thorough review of research on collaborative tagging. We showed that while collaborative tagging resembles traditional subject indexing, it offers much more flexibility and advantages because it encourages individual users to contribute their annotations in a collaborative manner. Studies in the literature were found to span a wide range of topics, including analyses of usage patterns of popular collaborative tagging systems, large scale clustering of tags and documents, investigations of the usefulness of tags in enhancing information retrieval on the Web, building recommendation systems based on collaborative tagging, and the generation of more structured metadata from folksonomies. While there was abundant research on both the characteristics and potentials of applications

of folksonomies, there is a lack of in-depth study of the mechanisms of how entities in a folksonomy acquire their semantics from the implicit associations generated by the tagging activities. This was the central motivation of this thesis.

We studied the notion of collective semantics in collaborative tagging by focusing on, one at a time, each of the three types of entity found in a folksonomy. We used the large amount of real-world data collected from Delicious as we described in Chapter 3. We first studied the semantics of tags in Chapter 4. Instead of referring to the meanings defined in a dictionary, we were interested in the social meanings of tags, the collective meanings given to the tags by the users when they use them. We compared different types of network representations of a folksonomy, and employed a network clustering algorithms on these networks to perform tag contextualisation. We found that networks that were constructed by explicitly taking user information into consideration were best in revealing the different contexts in which an ambiguous tag was used. This supported our hypothesis (Hypothesis 1) that networks that take into account the collective behaviour of the users are better in capturing the associations between different tags. We also presented in the same chapter the possibility of using the results of tag contextualisation in enhancing Web search by performing classification of search results. This chapter contributed significantly to understanding the semantics of tags in folksonomies.

In Chapter 5, we turned our attention to the users. It has been found in several studies that merely judging the expertise/trustworthiness of a user by how active he/she is can lead to mistaking malicious users such as spammers as good users. We proposed that such users can be uncovered by analysing their collective behaviour. We introduced the notion of implicit endorsement between users in collaborative tagging, and came up with two assumptions of experts: that there exists a mutual reinforcement relationship between the expertise of users and the quality of documents, and that experts are also discoverers of high quality documents. Based on these ideas, we proposed a graph-based algorithm, SPEAR, to rank users according to their expertise. Since there were no ground truths for evaluation, we came up with different models of expert users and spammers, and used these models to conduct a simulation-based evaluation of SPEAR. The results of our experiments supported our hypothesis (Hypothesis 2) that the qualities of the users can be inferred from their own behaviours and interactions. They also showed that our ideas are effective in reducing the negative influence of malicious users in a tagging system. We believe that the proposed algorithm, the models of users and the evaluation method all contributed to a better understanding of user

ranking on the Web.

Finally, we studied the semantics of documents in Chapter 6. One of the most important features of the Web is its ability to direct users from one document to another such that the users can discover useful and relevant information. Hence, we focused on the relations between documents in a folksonomy. We studied two different approaches to generating user-induced links from user tagging activities. One based on measuring similarity between tags of documents and one based on mining association rules in user collections. Our experiments showed that user-induced links could be considered as better recommendation links than existing hyperlinks. In other words, this result provided evidence that supports our hypothesis (Hypothesis 3) about implicit relations between documents that are generated by collaborative tagging: they are much more likely to lead users to useful and relevant information. In particular, we found that user-induced links based on rule mining of user preferences were more refined and of higher quality than those based on document similarity. Based on our findings, we proposed the use of user-induced links in predicting the tags of a document. Our experiments showed positive results, implying that user-induced links can be useful in document classification as well. Our study suggests that it is feasible to generate useful links from the collective user preferences found in a folksonomy. We believe this study contributes to a better understanding of relations between documents from the perspective of users.

While we have presented the three studies separately, it should be emphasised that they are actually closely related to each other and inter-dependent. In tag contextualisation, the number of spammers that could affect the result may have to be taken into consideration in some situations. Finding experts with respect to a topic represented by an ambiguous tag will in turn depend the results of tag contextualisation to single out the right users. Mining user-induced links may also benefit from the information that some users are more knowledgeable in a topic such that their preferences should be given more weights.

These three studies presented a thorough investigation of the notion of collective semantics in collaborative tagging. We demonstrated that the semantics—meanings, qualities and characteristics—of the entities involved in a folksonomy can be uncovered and understood by analysing the associations and network structures resulted from the collective tagging activities of the users. In particular, they showed that the analysis of user behaviours and interactions is crucial in the contextualisation of entities on the Web. We also demonstrated that these results can

benefit users by facilitating organisation and retrieval of information in a social setting on the Web.

The work described in this thesis fits into a larger framework for understanding the Web set out by the emerging interdisciplinary field of Web Science (Berners-Lee et al., 2006; Hendler et al., 2008). Web Science aims at studying both the social and technical aspects of the Web by bringing knowledge and ideas from such disciplines as computer science, psychology, economics and law. In particular, it emphasises the fact that the Web is not merely a technological product but also a social phenomenon that has grown so large that it deserves better understanding beyond its technical aspects. Collaborative tagging systems are like the Web in the way that they are all systems that allow a rather simple form of social interaction, which nevertheless eventually give rise to complex network structures among the tags, users and documents. Our work follows the line of thought of Web Science in that we believe and show that a better understanding of user behaviours allows us to understanding the semantics of the entities involved in collaborative tagging.

Clearly, the study of collective semantics is not completed with the conclusion of this thesis. We believe this thesis, while succeeding in answering some important questions, has opened up many possibilities for future research with respect to social interactions on the Web. In the remaining sections of this thesis, we will outline a number of future research directions, and discuss the outlook for research on collaborative tagging and in general the Social Web.

7.2 Future Research Directions

Throughout the course of this research, we came across a number of interesting questions about collective semantics in collaborative tagging and the Social Web. However, due to various constraints we were not able to further study these issues. We summarise in this section some issues that deserve further investigation and the possible methods of carrying out further research.

On a smaller scale, each of the three studies described in this thesis generate further questions about the subjects with which they were concerned. We therefore discuss each of these topics separately as follows.

7.2.1 On Tag Semantics

One issue in tag contextualisation we have not looked into is the granularity of the contexts discovered in the process. While the method we described in Chapter 4 offered satisfactory results, one may demand more information from the process of tag contextualisation. For example, even if we can identify two contexts in which the tag `sf` is used, namely 'San Francisco' and 'science fiction', it may also be desirable in some cases to identify some sub-contexts, such as 'travel information related to San Francisco' versus 'art studios in San Francisco'. This problem is actually related to the *resolution* of the network clustering algorithm used in the process (Fortunato and Barthelemy, 2007), i.e. how small a community can the algorithm detect within a given network. We believe that since there is no single correct answer to the contextualisation of a particular tag, there is no single best algorithm that can be used for the task. Instead, we may consider iterative algorithms such as proposed by Blondel et al. (2008). The resolution can be determined as needed at the time of execution. Tag contextualisation is important in enhancing the organisation and retrieval of documents in a folksonomy, and therefore this aspect deserves further investigation.

In addition, the semantics of tags are dynamic, not static. Some meanings may no longer be used by the users, while new meanings can keep appearing as new concepts emerge. Hence, we believe our study of the social meanings of tags should be extended to accommodate the temporal dynamics (Kleinberg, 2006) of tags. For example, can we detect when a certain meaning of a tag died away and when a new meaning of a tag emerged as a substantial number of users began to use it? This kind of analysis would be very useful in the context of information retrieval on the Social Web, because this will allow us to identify documents that are most relevant to the tag at the current time.

7.2.2 On User and Document Ranking

Concerning expertise ranking as discussed in Chapter 5, the proposed algorithm SPEAR can be further developed to accommodate more complex scenarios. Currently, SPEAR requires a pre-processing of the data by singling out documents that have been assigned the tag representing the topic we are interested in. One problem with such a requirement is that a topic can be represented by different tags. While it is true that we can filter documents by using a conjunction of multiple tags, in many cases we do not know which tags should be included, espe-

cially when there are usually many synonyms in a folksonomy (Niwa et al., 2007). While assessing the users' expertise in `javascript`, should we also consider their expertise in `programming` or `webdev`? This is not a trivial question and it deserves further investigation. A possible way of tackling this problem is to incorporate co-occurrence analysis of tags into SPEAR, and extend the reinforcement algorithm by accommodating users that have used highly related tags.

In addition, one more possible extension of the work on SPEAR is to investigate the document rankings it produces. One limitation of the current version of SPEAR is that it does not penalise users who tag a huge number of irrelevant documents, as tagging more documents will always increase a user's expertise score. However, a document's score will still be low if no other users tag them with the same tag(s). Hence, we believe a combination of document quality scores and user expertise scores will produce rankings that are more resistant to spamming activities.

7.2.3 On Implicit Relations between Documents

In our study of user-induced links we also focused on one tag at a time. Hence, a similar research question can be asked here: how can we take into account tags that are closely related or are synonymous? More generally, what we would like to see is a method that allows us to discover more sophisticated implicit relations between documents. For example, if many of the users who have assigned the tag `programming` to document d_1 would assign the tag `interesting` to document d_2, we may infer that document d_2 is interesting to users who like programming. This kind of algorithm will no doubt be very useful for understanding collective semantics in collaborative tagging and the Social Web in general, as more semantic relations between documents can be discovered. Moreover, we also believe that a user study that examines the usefulness of user-induced links would be very useful in confirming the value of links discovered in folksonomies.

In addition, we can extend this study of user-induced links to study the differences between user preferences and citation relationships between academic publications. For example, the collaborative tagging systems CiteULike and BibSonomy allow users to assign tags to publications. Would user-induced links be discovered between publications that do not cite each other? Or would user-induced links produce a path through the citation graph that effectively allows a student to learn about the topic of these publications? In this case, we may get a different picture of how publications are related to each other besides co-authorship and citation

relationship. In general, the methods mentioned in this thesis will be very useful for studying relations between documents from different perspectives.

On a larger scale, we expect to investigate how the methods of harvesting collective semantics discussed in this thesis can be adapted to handle a much larger amount of data. While we achieved promising results in our experiments, we are aware of the fact that we only focused on a single collaborative tagging system, from which we only harvested a relatively small set of data. The numerous collaborative tagging systems on the Web provide a lot of data on social interactions, but they also pose significant challenges to the processing of that data. In addition, this kind of data is not static but dynamic in the sense that its size keeps increasing as more users contribute more content to the Web. In other words, we need algorithms that can handle a huge amount of data and that can process data streams efficiently to reflect the most up-to-date picture.

There are two possible approaches to tackle these sorts of problem. A first approach is to consider parallel processing as a method to increase the amount of data that can be processed within a given period. An example of such an approach is the MapReduce framework introduced by Google for distributed computing of large data sets (Dean and Ghemawat, 2008). It is worthwhile to investigate how the algorithms considered in this thesis can be adapted and implemented using MapReduce or other similar frameworks to allow efficient processing of data on the Social Web. On the other hand, we can also consider employing data stream mining algorithms to perform the tasks described in this thesis (Gaber et al., 2005).

In summary, there are many possibilities that deserve further exploration. In the long run, we also look forward to moving beyond collaborative tagging and applying the techniques discussed in this thesis to other Social Web applications such as collaborative filtering sites and social networking sites to study collective semantics in a more general setting, and ultimately contribute to facilitating organisation, retrieval and sharing of information on the Web. In addition, for a better understanding of the interactions between online social networking and user behaviours, interdisciplinary directions along the line advocated by Web Science should also be pursued. In particular, it would be interesting to study how social theories can be applied to explain the interactions between users in different online systems, and some efforts in this direction can already be found in studies of social networking sites by Crandall et al. (2008), Danescu-Niculescu-Mizil et al. (2009) and Matsuo

and Yamamoto (2009).

7.3 The Future of Collaborative Tagging and the Social Web

Collaborative tagging is one of the prominent features of the Social Web. Its usefulness and potential are clearly shown in its ubiquity and popularity. From bookmarks to video clips, from photos to academic publications, from books on one's bookshelf at home to books in a university library, we only see increasing number of applications of tags on the Web. The Social Web is expanding rapidly, as forms of social interactions continue to evolve. For example, Twitter is currently very popular these days.[1] It is a micro-blogging service that allows users to read and write short messages for quicker social interactions on the Web and also on their mobile devices. Users using this service produce a huge amount of data everyday, which contain information about nearly every aspect of the users and their interests. This implies that we need better understanding of the dynamics on the Social Web, not only to facilitate sharing, organisation and retrieval of information, but also to understand the Web itself. For example, we may need better social network analysis to understand the dynamics of user behaviours underlying the Web, so that we can utilise the information to uncover useful collective semantics.

However, in terms of data portability, we see one significant limitation of tagging and in general the Social Web, which is that user-contributed data is usually confined within a single Web site such that data sets across different Web sites cannot be combined for better use. For example, we cannot use one tag to retrieve Web documents tagged in Delicious and photos tagged in Flickr at the same time. An academic publication saved on Bibsonomy is not associated with a Web document saved in Delicious, even though they may be of interest to the same group of users and are assigned the same set of tags. There are studies that attempt to analyse correlations between different folksonomies. For example, Szomszor et al. (2008b) study how user profiles in different folksonomies can be combined to produce a clearer picture of user interests. However, at the application level different Social Web applications remain isolated islands.

Given the merits of collaborative tagging and other Social Web applications, it

[1] Twitter: `http://www.twitter.com/`

would be desirable to have a universal infrastructure that promotes social interactions across different applications, and a standard that facilitates interoperability between social data in different domains. One of the possible solutions to these problems is how Semantic Web (Berners-Lee et al., 2001) technologies and Social Web applications can be combined. As for collaborative tagging, we have already discussed some studies that propose the use of ontologies to describe tagging activities on the Web. In a more general context, a prominent example is the Friend-Of-A-Friend (FOAF) Project (Brickley and Miller, 2007), which provides a set of vocabularies based on the Resource Description Framework (RDF) for users to define their social network on the Web. The Semantically-Interlinked Online Communities (SIOC) Project presented by Breslin et al. (2005) provides an ontology for describing social interactions on online community sites such as bulletin boards and mailing lists. Unlike commercial social networking sites such as Facebook and MySpace, an application that uses these Semantic Web technologies will provide data that can be exported in a standard format that can be integrated or reused with other data from another application. It can be envisioned that in order for user-contributed content to be more useful and more portable on the Web, Social Web applications will shift to a decentralised design in which data will not be stored within a single Web site but distributed in many different places and will be owned by the users themselves (Au Yeung et al., 2009d).

Bibliography

Karl Aberer, Philippe Cudré-Mauroux, Aris M. Ouksel, Tiziana Catarci, Mohand S. Hacid Arantza Illarramendi, Vipul Kashyap, Massimo Mecella, Eduardo Mena, and Erich J. Neuhold. Emergent semantics principles and issues. In Karl Aberer, Philippe Cudré-Mauroux, and Aris M. Ouksel, editors, *IFIP 2.6 Working Group on Data Semantics, 2004*, Zaragoza, 2004. Research Group of Distributed Information Systems (SID), University of Zaragoza.

Lada A. Adamic, Jun Zhang, Eytan Bakshy, and Mark S. Ackerman. Knowledge sharing and yahoo answers: everyone knows something. In *WWW '08: Proceeding of the 17th international conference on World Wide Web*, pages 665–674, New York, NY, USA, 2008. ACM.

Arun Kumar Agrahri, Divya Anand Thattandi Manickam, and John Riedl. Can people collaborate to improve the relevance of search results? In *RecSys '08: Proceedings of the 2008 ACM conference on Recommender systems*, pages 283–286, New York, NY, USA, 2008. ACM.

Rakesh Agrawal, Tomasz Imieliński, and Arun Swami. Mining association rules between sets of items in large databases. In *SIGMOD '93: Proceedings of the 1993 ACM SIGMOD international conference on Management of data*, pages 207–216, New York, NY, USA, 1993. ACM.

Luis Von Ahn, Manuel Blum, and John Langford. Captcha: Using hard ai problems for security. In *In Proceedings of Eurocrypt*, pages 294–311. Springer-Verlag, 2003.

Réka Albert, Hawoong Jeong, and Albert-László Barabási. Diameter of the world-wide web. *Nature*, 401:130–131, 1999.

Ching Man Au Yeung, Nicholas Gibbins, and Nigel Shadbolt. Mutual contextualization in tripartite graphs of folksonomies. In *ISWC 2007: The 6th International Semantic Web Conference, Pusan, South Korea, 11-15 November*, pages 966–970. Springer, 2007a.

Ching Man Au Yeung, Nicholas Gibbins, and Nigel Shadbolt. Tag meaning disambiguation through analysis of tripartite structure of folksonomies. In *The 2007 IEEE / WIC / ACM International Conference on Web Intelligence and Intelligent Agent Technology - Workshops, Silicon Valley, CA, USA, 2-5 November*, pages 3–6, 2007b.

Ching Man Au Yeung, Nicholas Gibbins, and Nigel Shadbolt. Understanding the semantics of ambiguous tags in folksonomies. In *The International Workshop on Emergent Semantics and Ontology Evolution (ESOE2007) at ISWC/ASWC 2007, Pusan, South Korea, 11-15 November*, pages 108–121, 2007c.

Ching Man Au Yeung, Nicholas Gibbins, and Nigel Shadbolt. Collective user behaviour and tag contextualisation in folksonomies. In *The 2008 IEEE / WIC / ACM International Conference on Web Intelligence and Intelligent Agent Technology - Workshops, Sydney, Australia, 9-12 December*, pages 659–662. IEEE Computer Society Press, 2008a.

Ching Man Au Yeung, Nicholas Gibbins, and Nigel Shadbolt. Discovering and modelling multiple interests of users in collaborative tagging systems. In *The 2008 IEEE / WIC / ACM International Conference on Web Intelligence and Intelligent Agent Technology - Workshops, Sydney, Australia, 9-12 December*, pages 115–118. IEEE Computer Society Press, 2008b.

Ching Man Au Yeung, Nicholas Gibbins, and Nigel Shadbolt. A k-nearest-neighbour method for classifying web search results with data in folksonomies. In *Proceedings of the 2008 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology, Sydney, Australia, 9-12 December*, pages 70–76. IEEE Computer Society Press, 2008c.

Ching Man Au Yeung, Nicholas Gibbins, and Nigel Shadbolt. A study of user profile generation from folksonomies. In *Proceedings of the Workshop on Social Web and Knowledge Management (SWKM2008) at WWW2008, Beijing, China, 21-25 April*, pages 1–8, 2008d.

Ching Man Au Yeung, Nicholas Gibbins, and Nigel Shadbolt. Web search disambiguation by collaborative tagging. In *Proceedings of the Workshop on Exploiting Semantic Annotations in Information Retrieval (ESAIR 2008), co-located with ECIR 2008, Glasgow, United Kingdom, 31 March 2008*, pages 48–61, 2008e.

Ching Man Au Yeung, Nicholas Gibbins, and Nigel Shadbolt. Contextualising tags in collaborative tagging systems. In *Proceedings of the 20th ACM Conference*

on Hypertext and Hypermedia, 29 June - 1 July, 2009, Turino, Italy, pages 251–260. ACM, 2009a.

Ching Man Au Yeung, Nicholas Gibbins, and Nigel Shadbolt. Multiple interests of users in collaborative tagging systems. In Ricardo Baeza-Yates and Irwin King, editors, *Weaving Services and People on the World Wide Web*, pages 255–274. Springer, 2009b.

Ching Man Au Yeung, Nicholas Gibbins, and Nigel Shadbolt. User-induced links in collaborative tagging systems. In *Proceedings of the 18th ACM Conference on Information and Knowledge Management, 2-6 November, 2009, Hong Kong.* ACM, 2009c.

Ching Man Au Yeung, Ilaria Liccardi, Kanghao Lu, Oshani Seneviratne, and Tim Berners-Lee. Decentralization: The future of online social networking. W3C Workshop on the Future of Social Networking, 15-16 January 2009, Barcelona, Spain, 2009d.

Ching Man Au Yeung, Michael G. Noll, Nicholas Gibbins, Christoph Meinel, and Nigel Shadbolt. On measuring expertise in collaborative tagging systems. In *WebSci'09: Web Science Conference 2009 - Society On-Line*, 2009e.

Soren Auer, Chris Bizer, Georgi Kobilarov, Jens Lehmann, Richard Cyganiak, and Zachary Ives. Dbpedia: A nucleus for a web of open data. In *Proceedings of the 6th International Semantic Web Conference (ISWC)*, volume 4825 of *LNCS*, pages 722–735. Springer, 2008.

Shenghua Bao, Guirong Xue, Xiaoyuan Wu, Yong Yu, Ben Fei, and Zhong Su. Optimizing web search using social annotations. In *WWW '07: Proceedings of the 16th international conference on World Wide Web*, pages 501–510, New York, NY, USA, 2007. ACM.

Albert-Laszlo Barabasi and Reka Albert. Emergence of scaling in random networks. *Science*, 286:509, 1999.

Albert-László Barabási, Réka Albert, and Hawoong Jeong. Scale-free characteristics of random networks: The topology of the world wide web. *Physica A*, 281: 69–77, 2000.

Luca Becchetti, Carlos Castillo, Debora Donato, Ricardo Baeza-YATES, and Stefano Leonardi. Link analysis for web spam detection. *ACM Trans. Web*, 2(1): 1–42, 2008.

Grigory Begelman, Philipp Keller, and Frank Smadja. Automated tag clustering: Improving search and exploration in the tag space. In *Collaborative Web Tagging Workshop at WWW2006, Edinburgh, Scotland*, 2006.

Pavel Berkhin. Survey of clustering data mining techniques. Technical report, Accrue Software, 2002.

Tim Berners-Lee, Wendy Hall, James A. Hendler, Kieron O'Hara, Nigel Shadbolt, and Daniel J. Weitzner. A framework for web science. *Foundations and Trends in Web Science*, 1(1):1–130, 2006.

Tim Berners-Lee, James Hendler, and Ora Lassila. The semantic web. *Scientific American*, 284(5):34–43, May 2001.

Howard Besser. Image databases: The first decade, the present, and the future. In P. Bryan Heydorn and Beth Sandore, editors, *Digital Image Access & Retrieval*, pages 11–28. University of Illinois, Urbana, IL, USA, 1997.

Kerstin Bischoff, Claudiu S. Firan, Wolfgang Nejdl, and Raluca Paiu. Can all tags be used for search? In *CIKM '08: Proceeding of the 17th ACM conference on Information and knowledge management*, pages 193–202, New York, NY, USA, 2008. ACM.

Vincent D. Blondel, Jean-Loup Guillaume, Renaud Lambiotte, and Etienne Lefebvre. Fast unfolding of community hierarchies in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, 2008.

Phillip Bonacich. Factoring and weighting approaches to status scores and clique detection. *Journal of Mathematical Sociology*, pages 113–120, 1972.

John G. Breslin, Andreas Harth, Uldis Bojars, and Stefan Decker. Towards semantically-interlinked online communities. In *Proceedings of the Second European Semantic Web Conference, ESWC 2005, Heraklion, Crete, Greece, May 29 - June 1, 2005*, volume 3532 of *LNCS*, pages 500–514. Springer, 2005.

Dan Brickley and Libby Miller. FOAF vocabulary specification. `http://xmlns.com/foaf/spec/`, 2007.

Sergey Brin and Lawrence Page. The anatomy of a large-scale hypertextual web search engine. In *Seventh International World-Wide Web Conference (WWW 1998)*, 1998.

Andrei Broder, Ravi Kumar, Farzin Maghoul, Prabhakar Raghavan, Sridhar Rajagopalan, Raymie Stata, Andrew Tomkins, and Janet Wiener. Graph structure in the web. *Computer Networks*, 33(1-6):309–320, 2000.

Christopher H. Brooks and Nancy Montanez. Improved annotation of the blogosphere via autotagging and hierarchical clustering. In *WWW '06: Proceedings of the 15th international conference on World Wide Web*, pages 625–632, New York, NY, USA, 2006. ACM Press.

Pauline Brown, Rob Hidderley, Hugh Griffin, and Sarah Rollason. The democratic indexing of images. *New Review of Hypermedia and Multimedia: Applications and Research*, 2:107–120, 1996.

Adriana Budura, Sebastian Michel, Philippe Cudre-Mauroux, and Karl Aberer. Neighborhood-based tag prediction. In *Proceedings of the 6th Annual European Semantic Web Conference*, 2009.

Vannevar Bush. As we may think. *The Atlantic Monthly*, July 1945.

Ciro Cattuto, Andrea Baldassarri, Vito D. P. Servedio, and Vittorio Loreto. Emergent community structure in social tagging systems. In *Proceedings of the European Conference on Complex Systems*, Dresden, Germany, October 2007.

Ciro Cattuto, Dominik Benz, Andreas Hotho, and Gerd Stumme. Semantic analysis of tag similarity measures in collaborative tagging systems. In *Proceedings of the 3rd Workshop on Ontology Learning and Population (OLP3)*, Patras, Greece, July 2008a.

Ciro Cattuto, Dominik Benz, Andreas Hotho, and Gerd Stumme. Semantic grounding of tag relatedness in social bookmarking systems. In Amith Sheth et al., editor, *The Semantic Web - ISWC 2008, Proc.Intl. Semantic Web Conference 2008*, volume 5318 of *LNAI*, pages 615–631, Heidelberg, 2008b. Springer.

Ciro Catutto, Christoph Schmitz, Andrea Baldassarri, Vito D. P. Servedio, Vittorio Loreto, , Andreas Hotho, Miranda Grahl, and Gerd Stumme. Network properties of folksonomies. *AI Communications Journal, Special Issue on "Network Analysis in Natural Sciences and Engineering"*, 2007.

Chandra Chekuri, Michael H. Goldwasser, Prabhakar Raghavan, and Eli Upfal. Web search using automated classification. In *Proceedings of the Sixth International World Wide Web Conference, Santa Clara, California, April, 2007*, 2007.

Liren Chen and Katia Sycara. Webmate: a personal agent for browsing and searching. In *AGENTS '98: Proceedings of the second international conference on Autonomous agents*, pages 132–139, New York, NY, USA, 1998. ACM.

Ed H. Chi. The social web: Research and opportunities. *Computer*, 41(9):88–91, 2008.

Ed H. Chi and Todd Mytkowicz. Understanding the efficiency of social tagging systems using information theory. In *HT '08: Proceedings of the nineteenth ACM conference on Hypertext and hypermedia*, pages 81–88, New York, NY, USA, 2008. ACM.

Michelene T. H. Chi. Two approaches to the study of experts' characteristics. In *The Cambridge Handbook of Expertise and Expert Performance*, pages 21–30. Cambridge University Press, New York, NY, USA, 2006.

Kenneth Ward Church and Patrick Hanks. Word association norms, mutual information, and lexicography. *Computational Linguistics*, 16(1):22–29, March 1990.

Maarten Clements, Arjen P. de Vries, and Marcel J. T. Reinders. Detecting synonyms in social tagging systems to improve content retrieval. In *SIGIR '08: Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*, pages 739–740, New York, NY, USA, 2008. ACM.

David Crandall, Dan Cosley, Daniel Huttenlocher, Jon Kleinberg, and Siddharth Suri. Feedback effects between similarity and social influence in online communities. In *KDD '08: Proceeding of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 160–168, New York, NY, USA, 2008. ACM.

Walt Crawford. Folksonomy and dichotomy. *Cites & Insights*, 6(4), 2006.

Douglass R. Cutting, David R. Karger, Jan O. Pedersen, and John W. Tukey. Scatter/gather: a cluster-based approach to browsing large document collections. In *SIGIR '92: Proc. of the 15th annual international ACM SIGIR conference on Research and development in information retrieval, Copenhagen, Denmark*, pages 318–329. ACM, 1992.

Cristian Danescu-Niculescu-Mizil, Gueorgi Kossinets, Jon Kleinberg, and Lillian Lee. How opinions are received by online communities: a case study on ama-

zon.com helpfulness votes. In *WWW '09: Proceedings of the 18th international conference on World wide web*, pages 141–150, New York, NY, USA, 2009. ACM.

Leon Danon, Jordi Duch, Albert Diaz-Guilera, and Alex Arenas. Comparing community structure identification. *Journal of Statistical Mechanics: Theory and Experiment*, 2005.

Wouter de Nooy, Andrej Mrvar, and Vladimir Batagelj. *Exploratory Social Network Analysis with Pajek*. Cambridge University Press, Cambridge, UK, 2005.

Jeffrey Dean and Sanjay Ghemawat. Mapreduce: simplified data processing on large clusters. *Commun. ACM*, 51(1):107–113, 2008.

Scott C. Deerwester, Susan T. Dumais, Thomas K. Landauer, George W. Furnas, and Richard A. Harshman. Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, 41(6):391–407, 1990.

Jörg Diederich and Tereza Iofciu. Finding communities of practice from user profiles based on folksonomies. In *Proceedings of the 1st International Workshop on Building Technology Enhanced Learning solutions for Communities of Practice*, 2006.

S. T. Dumais, G. W. Furnas, T. K. Landauer, S. Deerwester, and R. Harshman. Using latent semantic analysis to improve access to textual information. In *CHI '88: Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 281–285, New York, NY, USA, 1988. ACM.

Ayman Farahat, Thomas LoFaro, Joel C. Miller, Gregory Rae, and Lesley A. Ward. Authority rankings from hits, pagerank, and salsa: Existence, uniqueness, and effect of initialization. *SIAM J. Sci. Comput.*, 27(4):1181–1201, 2006.

Paul J. Feltovich, Michael J. Prietula, and K. Anders Ericsson. Studies of expertise from psychological perspectives. In *The Cambridge Handbook of Expertise and Expert Performance*, pages 41–68. Cambridge University Press, New York, NY, USA, 2006.

Raya Fidel. User-centered indexing. *Journal of the American Society for Information Science*, 45:572–576, 1994.

Gary William Flake, Steve Lawrence, C. Lee Giles, and Frans M. Coetzee. Self-organization and identification of web communities. *Computer*, 35(3):66–71, 2002.

Santo Fortunato and Marc Barthelemy. Resolution limit in community detection. *Proceedings of the National Academy of Sciences of the United States of America*, 104:36, 2007.

Andrew M. Fountain, Wendy Hall, Ian Heath, and Hugh C. Davis. MICROCOSM: an open model for hypermedia with dynamic linking. In *Hypertext: concepts, systems and applications*, pages 298–311. Cambridge University Press, New York, NY, USA, 1992. ISBN 0-521-40517-3.

G. W. Furnas, T. K. Landauer, L. M. Gomez, and S. T. Dumais. The vocabulary problem in human-system communication. *Commun. ACM*, 30(11):964–971, 1987.

George W. Furnas, Caterina Fake, Luis von Ahn, Joshua Schachter, Scott Golder, Kevin Fox, Marc Davis, Cameron Marlow, and Mor Naaman. Why do tagging systems work? In *CHI '06 extended abstracts on Human factors in computing systems*, pages 36–39, New York, NY, USA, 2006. ACM Press.

Mohamed Medhat Gaber, Arkady Zaslavsky, and Shonali Krishnaswamy. Mining data streams: a review. *SIGMOD Rec.*, 34(2):18–26, 2005.

Anestis Gkanogiannis and Theodore Kalamboukis. A novel supervised learning algorithm and its use for spam detection in social bookmarking systems. In *Proceedings of ECML PKDD Discovery Challenge Workshop, collocated with ECML/PKDD 2008*, 2008.

Daniela Godoy and Analia Amandi. User profiling in personal information agents: a survey. *Knowl. Eng. Rev.*, 20(4):329–361, 2005.

Scott Golder and Bernardo A. Huberman. Usage patterns of collaborative tagging systems. *Journal of Information Science*, 32(2):198–208, April 2006.

Thomas Gruber. Folksonomy of ontology: A mash-up of apples and oranges. *International Journal on Semantic Web and Information Systems*, 3(2), 2007.

Roger Guimera and Luis A. Nunes Amaral. Functional cartography of complex metabolic networks. *Nature*, 433:895, 2005.

A. Gulli and A. Signorini. The indexable web is more than 11.5 billion pages. In *WWW '05: Special interest tracks and posters of the 14th international conference on World Wide Web*, pages 902–903, New York, NY, USA, 2005. ACM.

Marieke Guy and Emma Tonkin. Folksonomies – tidying up tags? *D-Lib Magazine*, 12(1), 2006.

Harry Halpin, Valentin Robu, and Hana Shepherd. The complex dynamics of collaborative tagging. In *WWW '07: Proceedings of the 16th international conference on World Wide Web*, pages 211–220, New York, NY, USA, 2007. ACM.

Tony Hammond, Timo Hannay, Ben Lund, and Joanna Scott. Social bookmarking tools (i): A general review. *D-Lib Magazine*, 11(4), April 2005.

Peter Hannappel, Reinhold Klapsing, and Gustaf Neumann. MSEEC - a multi search engine with multiple clustering. In Mehdi KhosrowPour, editor, *Managing Information Technology Resources in Organizations in the Next Millennium: Proceedings of the 10th Information Resources Management Association International Conference*. Idea Group Publishing, 1999.

James A. Hendler, Nigel Shadbolt, Wendy Hall, Tim Berners-Lee, and Daniel J. Weitzner. Web science: an interdisciplinary approach to understanding the web. *Commun. ACM*, 51(7):60–69, 2008.

Monika Henzinger. Hyperlink analysis on the world wide web. In *HYPERTEXT '05: Proceedings of the sixteenth ACM conference on Hypertext and hypermedia*, pages 1–3, New York, NY, USA, 2005. ACM.

Jonathan L. Herlocker, Joseph A. Konstan, Loren G. Terveen, and John T. Riedl. Evaluating collaborative filtering recommender systems. *ACM Trans. Inf. Syst.*, 22(1):5–53, 2004.

Sven Herschel, Ralf Heese, Jens Bleiholder, and Christian Czekay. An architecture for emergent semantics. *Journal on Data Semantics XI*, pages 213–234, 2008.

Francis Heylighen. Mining associative meanings from the web: from word disambiguation to the global brain. In *Proceedings of Trends in Special Language and Language Technology*, pages 15–44. Standaard Publishers, 2001.

Paul Heymann and Hector Garcia-Molina. Collaborative creation of communal hierarchical taxonomies in social tagging systems. Technical Report 2006-10, Stanford University, April 2006.

Paul Heymann, Georgia Koutrika, and Hector Garcia-Molina. Fighting spam on social web sites: A survey of approaches and future challenges. *IEEE Internet Computing*, 11(6):36–45, 2007.

Paul Heymann, Georgia Koutrika, and Hector Garcia-Molina. Can social bookmarking improve web search? In *WSDM '08: Proceedings of the international*

conference on Web search and web data mining, pages 195–206, New York, NY, USA, 2008a. ACM. ISBN 978-1-59593-927-9.

Paul Heymann, Daniel Ramage, and Hector Garcia-Molina. Social tag prediction. In *SIGIR '08: Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*, pages 531–538, New York, NY, USA, 2008b. ACM.

Petter Holme, Mikael Huss, and Hawoong Jeong. Subnetwork hierarchies of biochemical pathways. *Bioinformatics*, 19:532, 2003.

Robert S. Hooper. *Indexer consistency tests: Origin, measurements, results and utilization*. IBM Corporation, 1965.

Andreas Hotho, Robert Jäschke, Christoph Schmitz, and Gerd Stumme. Information retrieval in folksonomies: Search and ranking. In York Sure and John Domingue, editors, *The Semantic Web: Research and Applications*, volume 4011 of *Lecture Notes of Computer Science*, pages 411–426. Springer, June 2006a.

Andreas Hotho, Robert Jäschke, Christoph Schmitz, and Gerd Stumme. Trend detection in folksonomies. In Yannis S. Avrithis, Yiannis Kompatsiaris, Steffen Staab, and Noel E. O'Connor, editors, *Proceedings of the First International Conference on Semantics And Digital Media Technology (SAMT)*, volume 4306 of *LNCS*, pages 56–70. Springer, 12 2006b.

Nancy Ide and Jean Veronis. Word sense disambiguation: The state of the art. *Computational Linguistics*, 24:1–40, 1998.

iProspect. Search engine user behaviour study, 2006.

Alex Iskold. The new face of amazon - tags, ajax, plogs & wikis. `http://www.readwriteweb.com/archives/amazon_tags_ajax_plogs_wikis.php`, January 2007.

Bernard J. Jansen, Amanda Spink, and Tefko Saracevic. Real life, real users, and real needs: a study and analysis of user queries on the web. *Inf. Process. Manage.*, 36(2):207–227, 2000.

Kalervo Järvelin and Jaana Kekäläinen. Cumulated gain-based evaluation of ir techniques. *ACM Trans. Inf. Syst.*, 20(4):422–446, 2002.

Thorsten Joachims. Optimizing search engines using clickthrough data. In *KDD '02: Proceedings of the eighth ACM SIGKDD international conference*

on Knowledge discovery and data mining, pages 133–142, New York, NY, USA, 2002. ACM.

Thorsten Joachims, Laura Granka, Bing Pan, Helene Hembrooke, and Geri Gay. Accurately interpreting clickthrough data as implicit feedback. In *SIGIR '05: Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 154–161, New York, NY, USA, 2005. ACM.

Ajita John and Dorée Seligmann. Collaborative tagging and expertise in the enterprise. In *WWW '06: Proceedings of the 15th international conference on World Wide Web*, New York, NY, USA, 2006. ACM Press.

Tomihisa Kamada and Satoru Kawai. An algorithm for drawing general undirected graphs. *Information Processing Letters*, 31(1):7–15, 1989.

Przemyslaw Kazienko and Marcin Pilarczyk. Hyperlink assessment based on web usage mining. In *HYPERTEXT '06: Proceedings of the seventeenth conference on Hypertext and hypermedia*, pages 85–88, New York, NY, USA, 2006. ACM.

Richard M. Keller, Shawn R. Wolfe, James R. Chen, Joshua L. Rabinowitz, and Nathalie Mathe. A bookmarking service for organizing and sharing urls. *Comput. Netw. ISDN Syst.*, 29(8-13):1103–1114, 1997.

Sue J. Ker and Jason S. Chang. A class-based approach to word alignment. *Comput. Linguist.*, 23(2):313–343, 1997.

Chanju Kim and Kyu-Baek Hwang. Naive bayes classifier learning with feature selection for spam detection in social bookmarking. In *Proceedings of ECML PKDD Discovery Challenge Workshop, collocated with ECML/PKDD 2008*, 2008.

Hak-Lae Kim, Simon Scerri, John Breslin, Stefan Decker, and Hong-Gee Kim. The state of the art in tag ontologies: A semantic model for tagging and folksonomies. In *International Conference on Dublin Core and Metadata Applications*, Berlin, Germany, 2008.

Hak-Lae Kim, Sung-Kwon Yang, Seung-Jae Song, John G. Breslin, and Hong-Gee Kim. Tag mediated society with scot ontology. In *The 5th Semantic Web Challenge, The 6th International Semantic Web Conference*, November 2007.

G. R. Kiss, C. Armstrong, R. Milroy, and J. Piper. An associative thesaurus of english and its computer analysis. In A. J. Aitken, R. W. Bailey, and

N. Hamilton-Smith, editors, *The Computer and Literary Studies*. University Press, Edinburgh, 1973.

Jon Kleinberg. Temporal dynamics of on-line information streams. In M. Garofalakis, J. Gehrke, and R. Rastogi, editors, *Data Stream Management: Processing High-Speed Data Streams*. Springer, 2006.

Jon Kleinberg, Ravi Kumar, Prabhakar Raghavan, Sridhar Rajagopalan, and Andrew Tomkins. The web as a graph: Measurements, models and methods. In *Proceedings of the International Conference on Combinatorics and Computing*, pages 1–18, Berlin, 1999. Springer.

Jon M. Kleinberg. Authoritative sources in a hyperlinked environment. *J. ACM*, 46(5):604–632, 1999.

Hyung Joon Kook. Profiling multiple domains of user interests and using them for personalized web support. In De-Shuang Huang, Xiao-Ping Zhang, and Guang-Bin Huang, editors, *Advances in Intelligent Computing, Proceedings of International Conference on Intelligent Computing (ICIC 2005), Part II, 23-26 August, 2005, Hefei, China*, pages 512–520, Secaucus, NJ, USA, 2005. Springer-Verlag New York, Inc.

Sherry Koshman, Amanda Spink, and Bernard J. Jansen. Web searching on the vivisimo search engine. *Journal of the American Society for Information Science and Technology*, 57(14):1875–1887, 2006.

Georgia Koutrika, Frans Adjie Effendi, Zolt'n Gyöngyi, Paul Heymann, and Hector Garcia-Molina. Combating spam in tagging systems: An evaluation. *ACM Trans. Web*, 2(4):1–34, 2008.

Beate Krause, Robert Jäschke, Andreas Hotho, and Gerd Stumme. Logsonomy - social information retrieval with logdata. In *HT '08: Proceedings of the nineteenth ACM conference on Hypertext and hypermedia*, pages 157–166, New York, NY, USA, 2008. ACM.

R. Krovetz and W. B. Croft. Word sense disambiguation using machine-readable dictionaries. *SIGIR Forum*, 23(SI):127–136, 1989.

Renaud Lambiotte and Marcel Ausloos. Collaborative tagging as a tripartite network. In *Proceedings of the International Conference on Computational Science*, volume 3993 of *LNCS*, page 1114. Springer-Verlag, 2006.

Frederick W. Lancaster. *Indexing and Abstracting in Theory and Practice*. University of Illinois Graduate School of Library and Information Science/Facet Publishing, Champaign, IL, USA, 3 edition, 2003.

Sara Shatford Layne. Subject access to art images. In Murtha Baca, editor, *Introduction to Art Image Access*. Getty Research Institute, Los Angeles, 2002.

Michael Lesk. Automatic sense disambiguation using machine readable dictionaries: how to tell a pine cone from an ice cream cone. In *SIGDOC '86: Proceedings of the 5th annual international conference on Systems documentation*, pages 24–26, New York, NY, USA, 1986. ACM.

Xin Li, Lei Guo, and Yihong Eric Zhao. Tag-based social interest discovery. In *WWW '08: Proceeding of the 17th international conference on World Wide Web*, pages 675–684, New York, NY, USA, 2008. ACM. ISBN 978-1-60558-085-2.

Ulrike Luxburg. A tutorial on spectral clustering. *Statistics and Computing*, 17 (4):395–416, 2007.

Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schütze. *Introduction to Information Retrieval*. Cambridge University Press, 2008.

Benjamin Markines, Heather Roinestad, and Filippo Menczer. Efficient assembly of social semantic networks. In *HT '08: Proceedings of the nineteenth ACM conference on Hypertext and hypermedia*, pages 149–156, New York, NY, USA, 2008. ACM.

Cameron Marlow, Mor Naaman, Danah Boyd, and Marc Davis. Ht06, tagging paper, taxonomy, flickr, academic article, to read. In *HYPERTEXT '06: Proceedings of the seventeenth conference on Hypertext and hypermedia*, pages 31–40, New York, NY, USA, 2006. ACM.

Adam Mathes. Folksonomies - cooperative classification and communication through shared metadata. `http://www.adammathes.com/academic/computer-mediated-communication/folksonomies.html`, December 2004.

Yutaka Matsuo and Hikaru Yamamoto. Community gravity: measuring bidirectional effects by trust and rating on online social networks. In *WWW '09: Proceedings of the 18th international conference on World wide web*, pages 751–760, New York, NY, USA, 2009. ACM.

Elke Michlmayr and Steve Cayzer. Learning user profiles from tagging data and leveraging them for personal(ized) information access. In *Proceedings of the*

Workshop on Tagging and Metadata for Social Information Organization, 16th International World Wide Web Conference (WWW2007), May 2007.

Peter Mika. Ontologies are us: A unified model of social networks and semantics. *Web Semantics*, 5(1):5–15, 2007.

David R. Millen, Jonathan Feinberg, and Bernard Kerr. Dogear: Social bookmarking in the enterprise. In *CHI '06: Proceedings of the SIGCHI conference on Human Factors in computing systems*, pages 111–120, New York, NY, USA, 2006. ACM. ISBN 1-59593-372-7.

George A. Miller. Wordnet: a lexical database for english. *Communications of the ACM*, 38(11):39–41, 1995.

Pierre-Alain Moëllic, Jean-Emmanuel Haugeard, and Guillaume Pitel. Image clustering based on a shared nearest neighbors approach for tagged collections. In *CIVR '08: Proceedings of the 2008 international conference on Content-based image and video retrieval*, pages 269–278, New York, NY, USA, 2008. ACM.

Miquel Montaner, Beatriz López, and Josep Lluís De La Rosa. A taxonomy of recommender agents on the internet. *Artif. Intell. Rev.*, 19(4):285–330, 2003.

D. K. Nelson, C. L. McEvoy, and T. A. Schreiber. The university of south florida word association, rhyme, and word fragment norms. `http://www.usf.edu/FreeAssociation/`, 1998.

Ted Nelson. *Literary Machines*. Mindful Press, Sausalito, California, 93.1 edition, 1993.

Nicolas Neubauer and Klaus Obermayer. Hyperincident connected components of tagging networks. In *Proceedings of 20th ACM Conference on Hypertext and Hypermedia*. ACM, 2009.

M. E. J. Newman. Analysis of weighted networks. *Physical Review E*, 70:056131, 2004a.

M. E. J. Newman. Fast algorithm for detecting community structure in networks. *Physical Review E*, 69:066133, 2004b.

M. E. J. Newman. Finding community structure in networks using the eigenvectors of matrices. *Physical Review E*, 74:036104, 2006.

M. E. J. Newman and M. Girvan. Finding and evaluating community structure in networks. *Physical Review E*, 69:026113, 2004.

Mark Newman, Albert-Laszlo Barabasi, and Duncan J. Watts. *The Structure and Dynamics of Networks: (Princeton Studies in Complexity)*. Princeton University Press, Princeton, NJ, USA, 2006. ISBN 0691113572.

Richard Newman. Tag ontology design. `http://www.holygoat.co.uk/projects/tags/`, 2004c.

NISO. Understanding metadata. `http://www.niso.org/publications/press/UnderstandingMetadata.pdf`, 2004.

Satoshi Niwa, Takuo Doi, and Shinichi Honiden. Web page recommender system based on folksonomy mining for itng'06 submissions. In *ITNG'06: Third International Conference on Information Technology: New Generations*, pages 388–393, 2006.

Satoshi Niwa, Takuo Doi, and Shinichi Honiden. Folksonomy tag organization method based on the tripartite graph analysis. In *IJCAI Workshop on Semantic Web for Collaborative Knowledge Acquisition*, January 2007.

Michael Noll and Christoph Meinel. Web search personalization via social bookmarking and tagging. In *Proceedings of the 6th International Semantic Web Conference and 2nd Asian Semantic Web Conference (ISWC/ASWC2007), Busan, South Korea*, pages 365–378, November 2007a.

Michael G. Noll, Ching Man Au Yeung, Nicholas Gibbins, Christoph Meinel, and Nigel Shadbolt. Telling experts from spammers: Expertise ranking in folksonomies. In *Proceedings of the 32nd Annual ACM SIGIR Conference, 19-23 July, 2009, Boston, MA, USA*, pages 612–619. ACM, 2009.

Michael G. Noll and Christoph Meinel. Authors vs. readers: a comparative study of document metadata and content in the www. In *DocEng '07: Proceedings of the 2007 ACM symposium on Document engineering*, pages 177–186, New York, NY, USA, 2007b. ACM.

Michael G. Noll and Christoph Meinel. Exploring social annotations for web document classification. In *SAC '08: Proceedings of the 2008 ACM symposium on Applied computing*, pages 2315–2320, New York, NY, USA, 2008a. ACM.

Michael G. Noll and Christoph Meinel. The metadata triumvirate: Social annotations, anchor texts and search queries. In *IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology*, pages 640–647, Los Alamitos, CA, USA, 2008b. IEEE Computer Society.

David S. PAlermo and James J. Jenkins. *Word Association Norms, Grade School Through Collect*. University of Minnesota Press, Minneapolis, MN, USA, 1964.

Alexandre Passant and Philippe Laublet. Meaning of a tag: A collaborative approach to bridge the gap between tagging and linked data. In *Proceedings of the WWW 2008 Workshop Linked Data on the Web (LDOW2008), Beijing, China*, 2008.

Elaine Peterson. Beneath the metadata: Some philosophical problems with folksonomy. *D-Lib Magazine*, 12(11), November 2006.

Steve Pinker. *The Stuff of Thought*. Viking Adult, 2008.

Raymond K. Pon, Alfonso F. Cardenas, David Buttler, and Terence Critchlow. Tracking multiple topics for finding interesting articles. In *KDD '07: Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 560–569, New York, NY, USA, 2007. ACM.

Emanuele Quintarelli. Folksonomies: power to the people. ISKO Italy-UniMIB meeting, 2005.

Filippo Radicchi, Claudio Castellano, Federico Cecconi, Vittorio Loreto, and Domenico Parisi. Defining and identifying communities in networks. *Proceedings of the National Academy Sciences of the United States of America*, 101: 2658, 2004.

Lee Rainie. 28% of online americans have used the internet to tag content. Technical report, Pew Internet and American Life Project, 2007.

David Resnick. Politics on the internet: the normalization of cyberspace. In *The Politics of Cyberspace*, pages 48–68. Routledge, New York, 1998.

Paul Resnick, Ko Kuwabara, Richard Zeckhauser, and Eric Friedman. Reputation systems. *Commun. ACM*, 43(12):45–48, 2000.

Paul Resnick and Richard Zeckhauser. Trust among strangers in internet transactions: Empirical analysis of ebay's reputation system. In Michael R. Baye, editor, *The Economics of the Internet and E-Commerce*, volume 11 of *Advances in Applied Microeconomics*, pages 127–157. Elsevier Science, Amsterdam, 2002.

John I. Saeed. *Semantics*. Wiley-Blackwell, 2003.

Timothy Salthouse. Expertise as the circumvention of human processing limitations. In K. Anders Ericsson, editor, *Toward a General Theory of Expertise*, pages 286–300. Cambridge University Press, New York, USA, 1991.

Gerard Salton and Chris Buckley. Term weighting approaches in automatic text retrieval. Technical report, Cornell University, Ithaca, NY, USA, 1987.

Christoph Schmitz, Andreas Hotho, Robert Ja"schke, and Gerd Stumme. Mining association rules in folksonomies. In V. Batagelj, H.-H. Bock, A. Ferligoj, and A. Ziberna, editors, *Data Science and Classification. Proceedings of the 10th IFCS Conf.*, Studies in Classification, Data Analysis, and Knowledge Organization, pages 261–270, Heidelberg, July 2006. Springer.

Patrick Schmitz. Inducing ontology from flickr tags. In *Collaborative Web Tagging Workshop at WWW2006*, Edinburgh, UK, 2006.

Hinrich Schütze. Automatic word sense discrimination. *Computational Linguistics*, 24(1):97–123, 1998.

Richard Segal and Jeffrey O. Kephart. Incremental learning in swiftfile. In *ICML '00: Proceedings of the Seventeenth International Conference on Machine Learning*, pages 863–870, San Francisco, CA, USA, 2000. Morgan Kaufmann Publishers Inc. ISBN 1-55860-707-2.

Shilad Sen, Shyong K. Lam, Al Mamunur Rashid, Dan Cosley, Dan Frankowski, Jeremy Osterhouse, F. Maxwell Harper, and John Riedl. Tagging, communities, vocabulary, evolution. In *CSCW '06: Proceedings of the 2006 20th anniversary conference on Computer supported cooperative work*, pages 181–190, New York, NY, USA, 2006. ACM Press.

Claude Elwood Shannon. *The Mathematical Theory of Communication*. University of Illinois Press, Urbana, Illinois, 1949.

Dou Shen, Jian-Tao Sun, Qiang Yang, and Zheng Chen. A comparison of implicit and explicit links for web page classification. In *WWW '06: Proceedings of the 15th international conference on World Wide Web*, pages 643–650, New York, NY, USA, 2006. ACM.

Kaikai Shen and Lide Wu. Folksonomy as a complex network. *The Computing Research Repository*, September 2005.

Andriy Shepitsen, Jonathan Gemmell, Bamshad Mobasher, and Robin Burke. Personalized recommendation in social tagging systems using hierarchical clustering. In *RecSys '08: Proceedings of the 2008 ACM conference on Recommender systems*, pages 259–266, New York, NY, USA, 2008. ACM.

Kei Shiratsuchi, Shinichiro Yoshii, and Masashi Furukawa. Finding unknown interests utilizing the wisdom of crowds in a social bookmark service. In *WI-IATW '06: Proceedings of the 2006 IEEE/WIC/ACM international conference on Web Intelligence and Intelligent Agent Technology*, pages 421–424, Washington, DC, USA, 2006. IEEE Computer Society.

Clay Shirky. Shirky: Ontology is overrated – categories, links, and tags. `http://shirky.com/writings/ontology_overrated.html`, 2005.

Craig Silverstein, Hannes Marais, Monika Henzinger, and Michael Moricz. Analysis of a very large web search engine query log. *SIGIR Forum*, 33(1):6–12, 1999.

Lucia Specia and Enrico Motta. Integrating folksonomies with the semantic web. In *ESWC '07: Proceedings of the 4th European conference on The Semantic Web*, pages 624–639, Berlin, Heidelberg, 2007. Springer-Verlag.

Steffen Staab. Emergent semantics. *IEEE Intelligent Systems*, 17(1):78–86, 2002.

Jerzy Stefanowski and Dawid Weiss. Carrot^2 and language properties in web search results clustering. In Ernestina Menasalvas Ruiz, Javier Segovia, and Piotr S. Szczepaniak, editors, *Proceedings of First International Atlantic Web Intelligence Conference, AWIC 2003, Madrid, Spain, May 5-6, 2003*, volume 2663 of *LNCS*, pages 240–249. Springer, 2003.

Fabian M. Suchanek, Gjergji Kasneci, and Gerhard Weikum. Yago - a core of semantic knowledge. In *Proceedings of the 16th International World Wide Web Conference*, 2007.

Fabian M. Suchanek, Milan Vojnovic, and Dinan Gunawardena. Social tags: meaning and suggestions. In *CIKM '08: Proceeding of the 17th ACM conference on Information and knowledge management*, pages 223–232, New York, NY, USA, 2008. ACM.

James Surowiecki. *The Wisdom of Crowds: Why the Many Are Smarter Than the Few and How Collective Wisdom Shapes Business, Economies, Societies and Nations.* Doubleday, 2004.

Martin N. Szomszor, Iván Cantador, and Harith Alani. Correlating user profiles from multiple folksonomies. In *HT '08: Proceedings of the nineteenth ACM conference on Hypertext and hypermedia*, pages 33–42, New York, NY, USA, 2008a. ACM.

Martin N. Szomszor, Iván Cantador, and Harith Alani. Correlating user profiles from multiple folksonomies. In *HT '08: Proceedings of the nineteenth ACM conference on Hypertext and hypermedia*, pages 33–42, New York, NY, USA, 2008b. ACM.

Tadeusz M. Szuba. *Computational Collective Intelligence*. John Wiley & Sons, Inc., New York, NY, USA, 2001.

Jennifer Trant. Social classification and folksonomy in art museums: Early data from the steve.museum tagger prototype. In Jonathan Furner and Joseph T. Tennis, editors, *Proceedings 17th Workshop of the American Society for Information Science and Technology Special Interest Group in Classification Research*, 2006.

Jennifer Trant and Bruce Wyman. Investigating social tagging and folksonomy in art museums with steve.museum. In *Collaborative Web Tagging Workshop at WWW2006, Edinburgh, Scotland*, 2006.

Céline Van Damme, Martin Hepp, and Katharina Siorpaes. Folksontology: An integrated approach for turning folksonomies into ontologies. In *Bridging the Gap between Semantic Web and Web2.0*, 2007.

C. J. van Rijsbergen. *Information Retrieval*. Butterworths, London, 1979.

Thomas Vander Wal. Explaining and showing broad and narrow folksonomies. `http://www.personalinfocloud.com/2005/02/explaining_and_.html`, 2005.

Sergei Vassilvitskii and Eric Brill. Using web-graph distance for relevance feedback in web search. In *SIGIR '06: Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 147–153, New York, NY, USA, 2006. ACM.

Miguel Villarroel, Pablo de la Fuente, Alberto Pedrero, Jesu's Vegas, and Joaqui'n Adiego. Obtaining feedback for indexing from highlighted text. *The Electronic Library*, 20(4):306–313, 2002.

Jakob Voss. Tagging, folksonomy & co - renaissance of manual indexing? In *Proceedings of the 10th International Symposium for Information Science*, 2007.

Jidong Wang, Zheng Chen, Li Tao, Wei-Ying Ma, and Liu Wenyin. Ranking user's relevance to a topic through link analysis on web logs. In *WIDM '02: Proceedings of the 4th international workshop on Web information and data management*, pages 49–54, New York, NY, USA, 2002. ACM. ISBN 1-58113-593-9.

Duncan J. Watts. *Small worlds: The Dynamics of Networks between Order and Randomness*. Princeton University Press, Princeton, NJ (USA), 1999.

David Weinberger. *Everything is Miscellaneous - The power of the new digital disorder*. Times Books, New York, 2007.

Aaron Weiss. The power of collective intelligence. *netWorker*, 9(3):16–23, 2005.

Robert Wetzker, Carsten Zimmermann, and Christian Bauckhage. Analyzing social bookmarking systems: A del.icio.us cookbook. In *Proceedings of Mining Social Data Workshop, collocated with ECAI 2008*, pages 26–30, 2008.

Ryen W. White, Mikhail Bilenko, and Silviu Cucerzan. Studying the use of popular destinations to enhance web search interaction. In *SIGIR '07: Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 159–166, New York, NY, USA, 2007. ACM.

Xian Wu, Lei Zhang, and Yong Yu. Exploring social annotations for the semantic web. In *WWW '06: Proceedings of the 15th international conference on World Wide Web*, pages 417–426, New York, NY, USA, 2006. ACM Press.

Zhichen Xu, Yun Fu, Jianchang Mao, and Difu Su. Towards the semantic web: Collaborative tag suggestions. In *Proceedings of the Collaborative Web Tagging Workshop, co-located with WWW 2006*, Edinburgh, Scotland, 2006.

Gui-Rong Xue, Hua-Jun Zeng, Zheng Chen, Wei-Ying Ma, Hong-Jiang Zhang, and Chao-Jun Lu. Implicit link analysis for small web search. In *SIGIR '03: Proceedings of the 26th Annual International ACM SIGIR conference*, pages 56–63, New York, NY, USA, 2003. ACM.

Yusuke Yanbe, Adam Jatowt, Satoshi Nakamura, and Katsumi Tanaka. Can social bookmarking enhance search in the web? In *JCDL '07: Proceedings of the 7th ACM/IEEE-CS joint conference on Digital libraries*, pages 107–116, New York, NY, USA, 2007. ACM.

Haifeng Yu, Michael Kaminsky, Phillip B. Gibbons, and Abraham Flaxman. Sybilguard: defending against sybil attacks via social networks. *SIGCOMM Comput. Commun. Rev.*, 36(4):267–278, 2006.

Oren Zamir and Oren Etzioni. Web document clustering: a feasibility demonstration. In *SIGIR '98: Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval*, pages 46–54, New York, NY, USA, 1998. ACM.

Hua-Jun Zeng, Qi-Cai He, Zheng Chen, Wei-Ying Ma, and Jinwen Ma. Learning to cluster web search results. In *SIGIR '04: Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 210–217, New York, NY, USA, 2004. ACM.

Xiaodan Zhang, Xiaohua Hu, and Xiaohua Zhou. A comparative evaluation of different link types on enhancing document clustering. In *SIGIR '08: Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*, pages 555–562, New York, NY, USA, 2008. ACM.

Ding Zhou, Sergey A. Orshanskiy, Hongyuan Zha, and C. Lee Giles. Co-ranking authors and documents in a heterogeneous network. In *ICDM '07: Proceedings of the 2007 Seventh IEEE International Conference on Data Mining*, pages 739–744, Washington, DC, USA, 2007a. IEEE Computer Society.

Mianwei Zhou, Shenghua Bao, Xian Wu, and Yong Yu. An unsupervised model for exploring hierarchical semantics from social annotations. In *The Semantic Web, 6th International Semantic Web Conference, 2nd Asian Semantic Web Conference, ISWC 2007 + ASWC 2007, Busan, Korea, November 11-15, 2007*, volume 4825 of *Lecture Notes in Computer Science*, pages 680–693. Springer, 2007b.