Exploiting Social Tagging Network for Web Mining and Search

A Thesis
Submitted to the Faculty
of
Drexel University
by
Caimei Lu
in partial fulfillment of the
requirements for the degree
of
Doctor of Philosophy
November 2011
ACKNOWLEDGEMENTS

First and foremost I offer my sincere gratitude to my advisor Xiaohua (Tony) Hu for his supervision, encouragement and support from the preliminary to the concluding level of my thesis. He showed me an excellent example of a successful researcher and professor. The energy, enthusiasm and joy he has for his research motivated all his students and the people around him, including me. I believe the experience of working with him over the past five years will continue to benefit me in my future career.

Dr. Jung-Ran Park deserves special thanks for being my co-advisor. I am much indebted to Dr. Park for her supervision at the initial stage of my Ph. D. study, and her valuable advice and funding support through the completion of this thesis.

My sincere thanks go to Dr. Il-Yeol Song who always kindly granted me his precious time for answering my questions and providing me useful advice on my research, teaching, and career development.

For this dissertation, I also would like to thank my committee members, Christopher Yang, Yuan An, and Jin Wen. I thank them for their constructive comments on this thesis and insightful questions in my thesis defense.

It is also a pleasure to pay tribute to the members of the research group supervised by Tony Hu. The group has been a source of good advice, collaboration and friendship. I benefited a lot from the outstanding work of Davis Zhou and Xiandan Zhang. The Dragon Toolkit developed by them was a very useful tool for my research experiments. Many thanks go to Xin Chen for research discussion and the pleasure working together in
the design of algorithms and development of research papers.

I am also grateful to all the fellow Ph.D. students I met at Drexel iSchool. I would like to thank Shanshan Ma, Palakorn Achananuparp, Ornsiri Thonggong, Ki Jung Lee, Jian Zhang, Sidath Gunawardena, Jia Huang, Zunyan Xiong, and Mi Zhang for being my great classmates, collaborators and friends and for all the fun we have had together in the last five years.

I gratefully acknowledge the funding sources that made my Ph.D. work possible. The research related to this thesis was supported in part by an IMLS Early Career Development Grant (2006-2010), NSF Career Grant (NSF IIS 0448023), and PA Department of Health Grant (No. 239667).

Lastly, I owe my deepest gratitude to my family. This thesis would not have been possible without their support. I wish to thank my parents for raising me as a responsible person and constantly supporting me in all my pursuits; My special gratitude goes to my elder brother Yonggang Lu, who had been a role model for me to follow since I was a teenager. I own him many thanks for showing me the joy of intellectual pursuit and generously helping me whenever I encountered difficulties in my study or life; And most of all, I wish to thank my loving, encouraging, and patient husband Haishan Zhao for his faithful support on my journey to the competition of this thesis.
# TABLE OF CONTENTS

Table of Contents .................................................................................................................. iii

List of Tables ......................................................................................................................... viii

List of Figures ......................................................................................................................... x

Abstract ..................................................................................................................................... xii

CHAPTER 1: Introduction ........................................................................................................ 1

1.1 The Potential of Social Tagging for Web Mining and Search ................................. 2

1.2 The Challenges of Utilizing Social Tagging for Web Mining and Search .......... 4

1.3 Research Questions and Framework ......................................................................... 7

1.3.1 The Effectiveness of Social Tags as Document Features ................................. 7

1.3.2 Tag-based Web Clustering ..................................................................................... 9

1.3.3 Personalized Search on Tagged Web ................................................................. 10

1.3.4 Topic Models of Tagged Web ............................................................................ 11

1.4 Data Sets for Experimentation .................................................................................... 13

1.5 The Organization of the Thesis ................................................................................... 13

CHAPTER 2: Literature Review ............................................................................................. 15

2.1 Usage Patterns and Semantics of Social Tags ......................................................... 15

2.1.1 Power Law Distribution ....................................................................................... 15

2.1.2 Stable Pattern in Social Tagging ......................................................................... 16

2.1.3 Tag Categories ..................................................................................................... 17

2.1.4 Meaningfulness of Social Tags ........................................................................... 18

2.2 Social Tags as Index terms ......................................................................................... 18
2.2.1 Social Tags and Professionally Created Metadata .................................................. 18
2.2.2 Social Tags, Full Text, and Query ........................................................................... 20
2.3 Tag-based Web Clustering ......................................................................................... 21
  2.3.1 Clustering on Extended Document Features ......................................................... 21
  2.3.2 Clustering on Social Tagging ................................................................................. 23
2.4 Personalized Search on Tagged Web ......................................................................... 24
  2.4.1 Personalized Search .............................................................................................. 24
  2.4.2 Personalized Search based on Social Tagging ....................................................... 25
2.5 Topic Models of Tagged Web ..................................................................................... 26
  2.5.1 Topic Analysis using Generative Models ............................................................... 26
  2.5.2 Topic Models of Tagged Web ................................................................................. 27
CHAPTER 3: A Comparison between Social Tags and Expert-assigned Subject Terms 30
  3.1 Introduction ............................................................................................................. 30
  3.2 Dataset .................................................................................................................... 33
  3.3 Analysis and results ............................................................................................... 35
    3.3.1 Overlapping vocabulary ..................................................................................... 36
    3.3.2 Frequent tags and LC subject headings ............................................................... 38
    3.3.3 Tags and LC subject headings annotating the same book .................................. 40
    3.3.4 Tags and LCSH subdivisions ............................................................................ 42
    3.3.5 Tags, LC subject headings and book titles ......................................................... 45
  3.4 Conclusion ............................................................................................................. 46
  3.5 Research Question Tested ...................................................................................... 48
CHAPTER 4: A Comparison of Social Tags, Author-Provided Metadata and Content

Words........................................................................................................................................49

4.1 Introduction..........................................................................................................................49

4.2 Dataset..................................................................................................................................52

4.3 Vocabulary Analysis.............................................................................................................54

4.4 Clustering Analysis..............................................................................................................57
   4.4.1 Clustering Standard and Quality Metrics.................................................................58
   4.4.2 Clustering Method and Results....................................................................................59

4.5 Conclusion............................................................................................................................63

4.6 Research Question Tested....................................................................................................64

CHAPTER 5: Clustering the Tagged Web .......................................................................................65

5.1 Introduction............................................................................................................................65

5.2 Tripartite Clustering Model ..................................................................................................67
   5.2.1 Model Formulation.........................................................................................................67
   5.2.2 Algorithm.......................................................................................................................71

5.3 Tag-based K-means and Link K-means..............................................................................75
   5.3.1 K-means.......................................................................................................................76
   5.3.2 Link K-means on Social Annotation............................................................................77

5.4 Experiments...........................................................................................................................80
   5.4.1 Dataset and Clustering Standard....................................................................................80
   5.4.2 Quality Metrics.............................................................................................................81
   5.4.3 Parameter Selection......................................................................................................81

5.5 Clustering Results..................................................................................................................83
LIST OF TABLES

Table 2.1 Mapping of four different classification schemas of social tags ............... 17
Table 3.1 Average frequency of shared terms and the whole set of tags and LCSH terms ........................................................................................................................................ 37
Table 3.2 Top 30 frequent tags and LC subject headings........................................... 39
Table 3.3 Number of book records with different percentages of LCSHs present in tags ......................................................................................................................................... 40
Table 3.4 Example terms assigned by users and experts to the same books and different books ........................................................................................................................................ 41
Table 3.5 Statistics of LCSH subfield terms................................................................. 43
Table 3.6 Top 10 LCSH subfield terms also used as tags ............................................ 43
Table 3.7 Number of book records which have at least one tag appearing in LCSH subfields ................................................................................................................................. 45
Table 4.1 30 most frequent tags and keyword terms ................................................. 57
Table 4.2 The clustering quality of K-means on different vector space models ......... 62
Table 5.1 Algorithm of Tripartite Clustering.............................................................. 75
Table 5.2 Quality of Tripartite Clustering for URLs ................................................ 84
Table 5.3 Top 10 Frequent Tags of 15 Tag Clusters .................................................... 85
Table 5.4 Clustering Quality of K-Means with Different Vector Space Models and Two Weighting Schemes .................................................................................................................. 87
Table 5.5 Clustering Quality of Link-Based K-Means Based on Different Link Types 88
Table 6.1 Searching methods used for experiments ................................................. 113
Table 6.2 MAP of different searching methods on Delicious dataset ................. 114
Table 6.3 Success of different searching methods on Delicious dataset .......... 114
Table 6.4  MAP of different searching methods on Bibsonomy dataset ................. 116
Table 6.5  Success of different searching methods on Bibsonomy dataset ............. 116
Table 7.1  Notations of Parameters .................................................................... 127
Table 7.2  A subset of discovered topics ............................................................ 134
Table 7.3  A subset of discovered perspectives ...................................................... 135
Table 7.4  Example Tags with three different values of $\lambda$ ............................ 136
LIST OF FIGURES

Figure 1.1 Illustration of the tripartite network of social tagging ........................................ 6

Figure 3.1 An example of a bibliographic MARC record .................................................. 34

Figure 3.2 Jaccard Index for Top frequent tags and LCSH terms ...................................... 38

Figure 4.1 Distribution of the 23,188 URLs over 14 ODP categories .................................. 53

Figure 4.2 The power law distribution of tag usage ......................................................... 54

Figure 4.3 The power law distribution of keyword usage .................................................. 54

Figure 5.1 The similarity between two documents r1 and r2 are not only decided by their direct links to users, but also affected by the cluster structure of users, which is further decided by the links from users to tags. ................................................................. 70

Figure 5.2 The distance between a document node and the centroid of a document cluster is influenced by the cluster structure of the tag nodes ......................................................... 73

Figure 5.3 The value of quality metrics for different values of $\alpha$ when $\beta=0.5$ and $\gamma=0.5$. ........................................................................................................................................ 82

Figure 5.4 The value of quality metrics for different values of $\beta$ when $\alpha=0.5$ and $\gamma=0.5$. ........................................................................................................................................ 83

Figure 5.5 The value of quality metrics for different values of $\gamma$ when $\alpha=0.5$ and $\beta=0.5$. ........................................................................................................................................ 83

Figure 6.1 An example of the tripartite social tagging hyper-graph .................................... 95

Figure 6.2 MAP@10 over different setting of $\alpha$ ............................................................. 108

Figure 6.3 Success@10 over different setting of $\alpha$ .......................................................... 109

Figure 6.4 The value of MAP@10 over different settings of $\lambda$ when word-based language model with Bayesian smoothing is incorporated into TripleQE to rank the candidate documents. .................................................................................. 111
Figure 6.5 The value of MAP@10 over different settings of $\lambda$ when tag-based language model with Bayesian smoothing is incorporated into TripleQE to rank the candidate documents. ................................................................. 111

Figure 7.1 The Topic-Perspective model for social tagging............................... 123

Figure 7.2 Graphical representation of (a) CI-LDA and (b) CorrLDA for modeling social annotations................................................................. 123

Figure 7.3 The perplexities over the iterations for five settings of topic number when perspective number $L=80$ ................................................................. 131

Figure 7.4 The perplexities over the iterations for five settings of perspective number when topic number $K=80$ ................................................................. 131

Figure 7.5 The perplexity results of CorrLDA, CI-LDA and TP-LDA (Topic-Perspective Model) for topic number $K=10, 20, 40, 80, 160$................................. 133
ABSTRACT
Exploiting Social Tagging Network for Web Mining and Search

Caimei Lu
Xiaohua Hu. Supervisor, Ph.D.

The rapidly growing social data created by users through Web 2.0 applications has intrigued active research in data mining and information retrieval (IR) community. The social tags created by users through social tagging system are one type of such social data. This thesis is dedicated to investigating whether and how the social tagging data can be utilized to enhance the performance of web mining and search methods.

First, in order to reveal whether social tags are effective document features which can be used to represent and index web documents, the author compares social tags with other type of index terms, including expert-created subject terms, author-provided keywords and description terms, as well as the content words of web documents. The results of the comparison studies show that social tags contain both high-quality index terms and subjective and personal terms. Besides, like author-provided keywords and description terms, social tags provide additional information beyond the content words of tagged web documents, but social tags are more effective than author-provided keywords and description terms as independent document features for web clustering.

The author further researches different approaches to improve web clustering performance by leveraging the social tagging data. The author proposes a novel clustering method called Tripartite Clustering which clusters web documents, users and
tags simultaneously based on the social tagging network. The author also investigates two other social tagging-based clustering approaches with K-means and Link K-means clustering methods. Experimental results show that all tag-based clustering methods can significantly improve the performance of content-based clustering. Compared to tag-based K-means and Link K-means, Tripartite Clustering achieves equivalent or better performance and produces more useful information.

The author also develops a novel personalized search framework based on a hypergraph model of social tagging. During the search process, the proposed framework combines three types of relations from the social tagging network for query expansion and ranking: the social relation among users, the semantic relation among tags, and the tripartite relation among users, tags and web documents. Experiments demonstrate that the proposed personalized search framework is more efficient and effective than baseline search methods.

Finally, the author proposes a novel topic model to simulate the generation of social tags and accordingly discover the topical structures of documents and users’ tagging perspectives. Experimental analysis shows that the proposed Topic-Perspective model has better generative ability than topic models proposed in existing literature. Besides, this model also generates more useful information about document topics, user perspectives, and the impact of document topics and user perspectives on tag generation.

In sum, this thesis not only reveals the potential value of social tags as document features and index terms of web documents, but also demonstrates how the social tagging network can be effectively utilized for web mining and search.
CHAPTER 1: INTRODUCTION

Web 2.0 based applications and systems have experienced explosive growth in the past several years. These applications and systems have reformed the way users manage, retrieve, share, and consume information. And more importantly, they provide the platforms for users to create new knowledge through online social activities such as tagging a web page, rating a book, commenting on a blog, answering a question, editing a wiki page, etc. These collaboratively created knowledge or social knowledge not only contain rich semantic information about the described web objects, but also provide a window for information providers to learn a user’s information interests and preferences. These days, the social knowledge created online users and their social activities have intrigued great interest among different research communities. Especially, in the data mining and information retrieval (IR) community, a serial of interesting research studies have been conducted to investigate how to improve the performance of traditional data mining and IR tasks by leveraging the social knowledge and network developed by online users. As one of the researchers inspired by the great potential of online social knowledge for data mining, in this thesis, the author aims to explore the information value of a typical type of social knowledge, i.e. the social annotations or tags created by online users in social tagging systems, and investigate its application for web mining and search.

Social tagging, also known as collaborative tagging, social classification, or social indexing, refers to the practice of users collaboratively creating and managing tags to annotate and categorize web resources. It has become one of the most popular Web 2.0
services since the first social bookmarking system Delicious was developed in 2003. By now, various social tagging applications have been developed for users to organize and retrieve web resources. The web resources tagged by users can be of any type and in any format, such as web pages (e.g. Delicious\(^1\)), videos (e.g. YouTube\(^2\)), photos (e.g. Flickr\(^3\)), academic papers (e.g. CiteULike\(^4\)), books (e.g. LibraryThings\(^5\)), music (e.g. Last.fm\(^6\)), and so on. Despite focusing on different types of resources, all the social tagging systems have the common purpose of helping users share, store, organize and search online resources that interest them. As more web resources get annotated every day, the annotations created by users and the network formed among users, tags and web resources have become valuable information sources for web mining and search.

1.1 The Potential of Social Tagging for Web Mining and Search

Social tagging can be exploited for web mining and search because of its several properties. Firstly, the tags created by users provide additional semantic information to the web resources. In this sense, social tags can be viewed as a special type of metadata which can be utilized to index and classify web resources. Vander Wal coined the term “folksonomy” (a portmanteau of the words folk and taxonomy) to describe the conceptual structure generated by social tagging systems (Wal 2007). Different from traditional taxonomies, which are created by professionals based on some strict standards and guidelines, the folksonomy is developed in an uncontrolled and spontaneous approach.

---

1 http://delicious.com/
2 http://www.youtube.com/
3 http://www.flickr.com/
4 http://www.citeulike.com/
5 http://www.librarything.com/
6 http://www.last.fm/
To a great extent, social tagging lowers the entrance threshold of metadata creation. Therefore, social tagging provides a cheaper and more efficient alternative to the traditional metadata creation approach. The tags applied to web resources can be utilized as additional document features to enrich the document representation and thus enhance the performance of data mining and retrieval in the web environment.

Secondly, social tagging adapts quickly to emerging vocabularies and topics. As the social culture and technology evolve, new terms and concepts continue to emerge in different domains. For instance, ‘Folksonomy’ itself is one of them. These new terms and concepts as well as their related resources can be absorbed into the social tagging system in real time, which is impossible for traditional ontologies or controlled vocabularies due to their strict and complex maintenance procedures. This indicates that the social tagging system can be utilized to retrieve resources related to emerging concepts and topics, and it may also be used to track new topics in different domains along the time dimension.

Thirdly, from the point of view of information retrieval, it is easier to achieve indexer-searcher consistency in social tagging systems, which is a generally agreed upon robust indicator of retrieval effectiveness (Furner 2007). In the information retrieval domain, effective indexing and retrieving means that indexers should be able to predict the terms that will be used by searchers to query the same resources. In social tagging systems, taggers are indexers and searchers at the same time. Therefore, the probability that indexers and searchers will agree on the subjects of a given resource and use the same combination of terms to express the given subjects would be higher in social tagging systems than in other indexing and metadata creation systems. This implies that better search queries can formulated based on the social tags applied to relevant
The last but not least property of social tagging that makes it valuable for web mining and search is that, it provides a channel to learn users’ information needs, preferences and interests, which are essential factors of personalized search and recommendation systems. Specifically, in a social tagging system, a user’s preference and interest can be characterized by the web resources she/he has ever annotated and the tags she/he has ever applied to those web resources. Although researchers have proposed solutions to extract users’ interests and preferences from various information sources such as users’ search and browsing history, these information sources are always unavailable to the public. For instance, users’ search, click-through and browsing history, recorded by specific search engine and information vendors, are normally unavailable to outsiders due to commercial and privacy causes. As opposed to these information sources, social tagging data have very high availability, because a main objective of users’ social tagging activity is to share information.

1.2 The Challenges of Utilizing Social Tagging for Web Mining and Search

Despite of the great potential of social tagging for web mining and search, the approaches of exploiting its potential value are not straightforward. There are challenges to overcome for social tagging being fully utilized to enhance traditional web mining and search tasks.

Firstly, the usage of non-words, polysemy, and synonymy, which usually cause semantic ambiguity and imprecision, is prevalent in social tagging. Suchanek et al. (2008) checked a sample of tags collected from Del.icio.us using two dictionaries: YAGO and WordNet. It was found that more than half of the sampled social tags are unknown to the dictionaries and most tags known to the dictionaries have more than one
meaning. A reason for this phenomenon is that, different from the controlled vocabularies of taxonomies and ontologies, folksonomy is developed in a totally uncontrolled environment. Users can basically use any terms to annotate a web resource in a social tagging system. Therefore, before utilizing social tagging data, it is necessary to investigate the quality and structures of social tags in general.

Besides, it has been shown that users tag web resources for purposes other than topical description. Golder and Huberman (2006) identify tags with seven different functions: (1) identify what (or who) it is about; (2) identify what it is; (3) identify who owns it; (4) refining categories; (5) identifying qualities or characteristics; (6) self-reference; (7) task organizing. Sen et al. (2006) summarize the seven tag functions proposed by Golder and Huberman into three categories: Factual tags, Subjective tags, and Personal tags. Intuitively, Factual tags are more closely related to resource content and extrinsic to the taggers, while the Subjective tags and Personal tags are less connected to resource topics and more influenced by users’ perspectives. Tags such as “to read”, “unread”, “funny”, “cool”, etc. are very common examples of personal and subjective tags. These subjective and personal tags may be useful for the tagger who created them, but may be valueless for other users. Comparatively, because factual tags are extrinsic to the taggers but closely related to the topics of web resources, they can be more effectively used as additional document features in general web mining and search tasks. Thus, it is desirable to distinguish factual tags from personal and subjective tags in social tagging-based applications. However, how to automatically classify the tags into factual tags and subjective and personal tags is non-straightforward and challenging.

Finally, different from previously well-studied bipartite and unipartite network
structures (such as the hyperlink network, author-document network, or document-word network), the social tagging network is composed by the tripartite relations among users, tags and web documents. Whenever a user annotates a web document with a tag, a tripartite relation is built among the user, the document and the tag.

Figure 1.1 Illustration of the tripartite network of social tagging. u, r, and t respectively denote user, resource and tag. Each time a user annotate a resource with a tag, a tripartite relationship is built among the user, the resource and the tag. This figure contains seven tripartite relationships.

Figure 1.1 illustrates the social tagging tripartite graph. The tripartite network of social tagging system contains valuable information for learning the topics of web resources, the semantics of tags, the interest of users, and their relations. However, modeling and mining tripartite graph is always demanding in terms of algorithm design and computational efficiency. In order to utilize the social tagging data, a common approach is to project the social tagging network into bipartite or unipartite networks, then apply existing models and algorithms for bipartite and unipartite relationships on social tagging data. For instance, a user-document-tag tripartite relation can be projected into three undirected bipartite relations: user-tag relation, user-document relation and
document-tag relation. Although it is always possible to project the tripartite network of social tagging into bipartite or unipartite networks, the projection inevitably causes information loss.

In this thesis, the author proposes different approaches to overcome these challenges in order to thoroughly investigate and utilize the value of social tagging in different web mining and search tasks.

1.3 Research Questions and Framework

Based on the understanding of the potentials and challenges of utilizing social tagging data for web mining and search, the author proposes and answers four research questions in this thesis.

Question 1: Are social tags effective document features which can be used to represent and index web documents in web mining and search applications?

Question 2: How to enhance web clustering based on the social tagging network?

Question 3: How to exploit the social tagging network for personalized web search?

Question 4: How to develop a topic model for the tagged web?

1.3.1 The Effectiveness of Social Tags as Document Features

There is no direct way of examining the effectiveness of social tags as document features or index terms. In this thesis, the author attains this research goal through three indirect approaches. First, the author compares user-created social tags with professional created index terms. Social tags have been recognized as a type of social metadata created by users through the web 2.0 applications (Smith-Yoshimura 2011). As an open and collaborative metadata creation approach, social tagging is much more efficient than the
traditional indexing approach based on metadata standards and controlled vocabularies. However, people are concerned with the quality of social tags as indexing terms because of its uncontrolled nature. In order to reveal the quality of social tags as index terms, an indirect but useful approach is to compare social tags with the subject terms created by information professionals in traditional information organization and retrieval environment. In this thesis (Chapter 3), the author compares social tags with the subject headings created by experts to investigate the differences and connections between these two types of metadata and thus indirectly proves the feasibility of using social tags as index terms for web documents.

Besides expert-assigned subject terms, the author also compares tags created by users with the metadata generated by the authors or creators of web documents, such as the keywords and description of a web page given by its authors (Chapter 4). Before the appearance of social tagging systems, authors were seen as an appropriate alternative of the experts as metadata creators, because it can help with the scalability problems in comparison to expert-created metadata, and authors are supposed to be capable of describing their work with precise terms because they are most intimate with their work (Greenberg et al. 2001). However, some researchers point out that authors are not necessarily any better at generating metadata than anyone else (Mathes 2008, Thomas and Griffin 1999), because authors and the intended users of their work may have very different understandings of their work and thus disagree with the terms that can be used to describe the work. In short, both expert-created metadata and author-created metadata has the problem: the intended and unintended eventual users of the information are disconnected from the indexing process (Mathes 2008). On the contrary, in social tagging
systems, taggers are at the same time indexers and information users. In this thesis, the author compares social tags with author-provided metadata in terms of their vocabulary and effectiveness for facilitating web clustering and discovery.

Finally, the author investigates whether social tags provide additional information value beyond the content of web resources. Most exiting text mining and information retrieval approaches are based on document content. For instance, most current search engines are founded on the full-text indexing technique. For social tags to be useful, they need to provide additional information value beyond document content. This indicates that the tags not only need to contain new information which cannot be derived from document content, the new information from social tags should also be valuable for identifying the documents. If social tags only comprises of terms copied from document content or the tags unfound in document content are noisy and subjective, they would contribute none value to web mining and search. In this thesis (Chapter 4 and 5), the author compares the social tags assigned to web pages with the content terms contained in the web pages based on a large real-world dataset. The author not only checks the difference between social tags and document content in term usage, but also compares the effectiveness of social tags and content terms in web mining tasks such as web clustering.

1.3.2 Tag-based Web Clustering

Clustering is an unsupervised data mining method which has been widely used in different areas. It can be used to automatically group thousands and millions of documents into a list of meaningful categories based on some document features. In traditional clustering algorithms, such as k-means, content terms are used as document features for clustering. Nevertheless, in the web environment, the content features of web
pages are sometimes missing, misleading and unrecognizable due to the lack of well-controlled authoring styles. For non-textual web documents, such as images and videos, the content features are not directly available. Therefore, it is desirable to exploit other information sources to enhance clustering effectiveness. For instance, hyperlinks connecting web documents have been used as important features for text clustering in previous research (Angelova and Siersdorfer 2006, Qi and Davison 2006). In this thesis (Chapter 5), the author view the social tags assigned to documents by online users as a type of rich information source for web mining. Social tag based clustering is compared to content-based clustering using different methods. Especially, based on the tripartite social tagging network, the author develops a novel clustering method (Tripartite Clustering) which is independent of document content and clusters documents, users and tags simultaneously.

1.3.3 Personalized Search on Tagged Web

Without proper information sources for user interests, the classical search engines have to assume that the relevance between a query and a document is only decided by the similarity of term matching. However, as pointed by Pitkow et al. (2002), relevance is actually relative for each user. Teevan et al. (2007) also conclude that users differ greatly in the search results they considered to be relevant to the same query. A solution to this problem is to develop personalized search systems by incorporating user-specific information into the search process. In this thesis, the author explores how to develop personalized search systems by leveraging the user information learned from the social tagging network. Compared to other user information sources adopted in most existing studies for personalized search, such as users’ search and browsing history saved by
specific search engines and information vendors, social tagging provides a more open and effective channel for learning users’ information needs and interests. In this thesis (Chapter 6), the author proposes an efficient and effective personalized search framework (TripleQE) based on the tripartite network of social tagging. The proposed search framework utilizes query expansion for personalization and ranking. Different from the traditional expansion methods which only expand the query keywords, this search framework also expands the queries in the user dimension. More importantly, it relies on the tripartite relationship from the social tagging network for query expansion. The relative efficiency and effectiveness of this framework is demonstrated by experiments on real-world data.

1.3.4 Topic Models of Tagged Web

In text mining and IR tasks, the choice of document features is a challenging problem. Traditional approaches usually represent documents based on word features. However, the word feature set is usually very large especially for large collections, which causes inefficiency in word-based applications. Therefore, dimensionality-reduction methods that can enable efficient processing of large collections are always desirable. A classic approach for dimensionality reduction is latent semantic analysis (LSA) (Deerwester et al. 1990). The effectiveness of LSA has been proven in a wide range of applications. However, it still has limitations in theoretical foundation and the ability to dealing with polysemy. Thus, topical models based on sound statistical foundation are developed. Topic models are probabilistic generative models which discover “latent topics” from a collection of documents. In a topic model, the representation of any documents is reduced to a probability distribution on a fixed set of topics. An early topic model called
Probabilistic Latent Semantic Indexing (PLSI) was proposed by Hoffman in 1999 (Hoffman 1999). Based on PLSI, Blei et al. (2003) proposes the Latent Dirichlet Allocation (LDA) model, which has become the most common topic model in use due to its solid theoretical foundation and promising performance. The LDA assumes that the probability distribution over words in a document can be represented as a mixture of topics, and each topic is a probability distribution over words. LDA has been used in many applications including clustering, classification and IR. Moreover, various extended LDA models have been proposed to exploit information other than document words for topic learning, such as extracted entities (Newman et al. 2006) and document links (Erosheva et al. 2004, Liu et al. 2009).

In this thesis, the author investigates how to model the tagged web based on LDA (Chapter 7). Based on the nature of social tagging process, the author develops a new topic model which not only models the generation of words, but also models the generation of tags. An important feature of this model is that, during the tag generation process, document topics and user perspectives together generate the social tags assigned to a resource, and each tag differs in the extent of depending on resource topics or user perspectives. Through this feature, the proposed model can be used to discover not only the latent topics from words and tags, but also the latent user perspectives. More importantly, it models tags with different functional purposes. It can be used to estimate whether a tag is more likely a factual tag created for topic description or a subjective/personal tag for opinion expression or self-reference. This feature makes the model useful for many applications.
1.4 Data Sets for Experimentation

For social tagging research, there is no standard dataset for experimentation. The common practice is to collect datasets from different real-word social tagging systems. In this thesis, the author prepares three social tagging datasets collected from three social tagging websites: Delicious, LibraryThing and Bibsonomy. Delicious is a social bookmarking website for storing, sharing and discovering web bookmarks. The data set from Delicious was crawled from its website during January 2009 and February 2009. The original dataset contains 3,246,424 posts created by 4784 users. Each post is a bookmark to a webpage annotated by a user with one or more tags. This is the primary dataset used in this dissertation research. The second dataset was crawled from LibraryThing, a social cataloging web service for users to tag, store and share book catalogs. The crawling process was conducted during October 2009. This data set is used for the comparison study of social tags and expert-assigned subject terms. Bibsonomy is also a social bookmarking web service, but it allows users tag both web pages and published articles. The Bibsonomy dataset is downloaded from its data dump page (http://www.kde.cs.uni-kassel.de/bibsonomy/dumps). The data dump created by Jan 01, 2011 was chosen for experimentation. The Bibsonomy dataset is primarily used to study personalized search based on social tagging network.

1.5 The Organization of the Thesis

The rest of the thesis is organized as follows: Chapter 2 reviews existing research studies on social tagging systems, including the usage patterns and semantic values of social tags, as well as the applications of social tagging data in document clustering, information
retrieval and topic modeling. In Chapter 3, an empirical study is conducted to compare social tags with expert-created subject terms. In Chapter 4, social tags are compared with author provided keywords and descriptions through an empirical study. Besides, both social tags and author-provided metadata are compared with content features at both vocabulary level and application level. Chapter 5 discusses possible approaches of utilizing social tagging for web clustering and proposes a novel clustering method called Tripartite Clustering, which clusters web documents, tags and users at the same time based on the tripartite network of social tagging. Chapter 6 introduces a personalized search framework based on the tripartite social tagging network. Chapter 7 develops a new topic model called Topic-Perspective model to simulate the generation process of words in web documents as well as the tags assigned to the web documents by users. Chapter 8 concludes the work of this thesis.
CHAPTER 2: LITERATURE REVIEW

Social tagging has received major attention in recent literature. Existing research on social tagging has been focused on the usage patterns and semantics of social tags, the indexing ability of social tags compared to other indexing approaches, and its application in text mining and information retrieval areas. In this chapter, we review the major literature on all these topics. Especially, for the third topic, the author focuses on the application of social tagging in clustering, information retrieval and web search, as well as topic modeling.

2.1 Usage Patterns and Semantics of Social Tags

Before studying the applications of social tags, it is worthwhile first investigating the usage patterns and semantics of social tags, the knowledge about which can help develop better application solutions. In this section, the author discusses some usage patterns and semantic values of social annotations identified in existing literature

2.1.1 Power Law Distribution

The most obvious usage pattern present in a social tagging system is power law. The frequency of both tagged objects (e.g. URL) and tags follows a power law distribution (Wu et al. 2006, Li et al. 2008). That means a very small portion of web objects are very frequently tagged, and most objects are only bookmarked by a few users. Likewise, a relatively small number of tags are highly used by users while most tags are only applied to a few objects. Previous research proved that this small portion of frequently used tags
usually comprises of tags with generic meanings (Xu et al. 2006, Golder and Huberman 2006). This indicates that it is possible to filter out the personal tags based on the tag frequency. Golder and Huberman (2006) also found that when both general tags and personal tags are used, “users have a strong bias toward using general tags first”. Besides the popularity of tagged objects and tags, the activity of users also follows the power law distribution (Li et al. 2008, Golder and Huberman 2006). However, Heymann et al. found that the top 10% of users only account for 56% of posts on Del.icio.us. Based on the tagging history of Delicious site, Harry Halpin et al. (2007) examined why and how the power law distribution of tag usage frequency was formed in a mature social tagging system along time.

2.1.2 Stable Pattern in Social Tagging

In a social tagging system, tagging is conducted in an uncontrolled environment and users are free to choose their own vocabularies to describe the objects. Therefore, one might expect that this would result in a chaotic tag collection. However, based on the dataset collected from Delicious, Golder and Huberman (2006) found that as a URL’s bookmarks accumulated, the portion of each tag applied to the URL followed a stable pattern. Especially, after about 100 bookmarks, the portion of each tag was nearly fixed. This pattern indicates that after a webpage was bookmarked by a certain number of users a consensus about the terms used for describing the webpage was formed. Interestingly, Suchanek et al. (2008) also found that the meaningfulness of a URL’s top popular tags increased significantly after the URL was tagged more than 100 times. These findings imply the potential value of popular social annotations as a reliable information source

2.1.3 Tag Categories

In order to better understand the semantic structure of folksonomy, some researchers analyzed different types of tags used in social tagging systems. For instance, Golder and Huberman (2006) classified tags into seven categories based on their functional purposes (see Table 1). Sen et al. (2006) summarized the seven categories of tags proposed by Golder and Huberman into three classes: Factual tags, Subjective tags, and Personal tags. Xu et al. (2006) proposed another taxonomy which contains five categories of tags. Based on Golder and Huberman’s schema, Bischoff et al. (2008) specified eight categories of tags, as shown in Table 1. Moreover, Bischoff et al. studied the distribution of tag types across three different social tagging systems: Delicious, Flicker and Last.fm. The results showed that the distributions of tag types vary among different social tagging systems. For instance, “Topic” is the most frequent tag type for Del.icio.us, while “Type” is most important for Last.fm. As for Flicker, “Topic” and “Location” are most popular tag types.

Table 2.1 Mapping of four different classification schemas of social tags

<table>
<thead>
<tr>
<th>(Bischoff et al. 2008)</th>
<th>(Golder et al. 2006)</th>
<th>(Xu et al. 2006)</th>
<th>(Sen et al. 2006)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Topic</td>
<td>What or who it is about</td>
<td>Content-based</td>
<td></td>
</tr>
<tr>
<td>Time</td>
<td>Refining categories</td>
<td>Context-based</td>
<td>Factual</td>
</tr>
<tr>
<td>Location</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Type</td>
<td>What it is</td>
<td>Attribute</td>
<td></td>
</tr>
<tr>
<td>Author/Owner</td>
<td>Who owns it</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Opinions/Qualities</td>
<td>Qualities and characteristics</td>
<td>Subjective</td>
<td>Subjective</td>
</tr>
<tr>
<td>Usage context</td>
<td>Task organization</td>
<td>Organizational</td>
<td>Personal</td>
</tr>
<tr>
<td>Self-reference</td>
<td>Self-reference</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
2.1.4 Meaningfulness of Social Tags

Since social tags are assigned by web users rather than trained professionals, people may wonder how correctly social tags describe the annotated objects. It has been assumed that non-words, polysemy and synonymy are prevalent in social tags. To verify this assumption, Suchanek et al. (2008) checked the meaning of a sample of tags collected from Delicious with two dictionaries YAGO and WordNet. It turned out that more than half of social tags are unknown to the dictionaries and most tags known to the dictionaries have more than one meaning. However, they also found that top popular tags of a webpage are usually semantically meaningful in the sense that these tags are registered to the dictionaries. Besides, they also found that a quarter of social tags were used as DMOZ category labels. In order to check whether social tags are meaningful and relevant descriptors of the annotated objects, Heymann et al. (2008) let a group of graduate students to evaluate a sample of posts collected from del.icio.us. The result of their study showed that most tags are relevant and objective. Based on a probabilistic generative model, Zhang et al. identified the conceptual structures of folksonomy and proved that a hierarchical relationship can be derived from the identified concepts in folksonomy (2006).

2.2 Social Tags as Index terms

2.2.1 Social Tags and Professionally Created Metadata

A common approach to evaluating the indexing ability of social tags is to compare them with the metadata created by professionals. Several studies have been done to compare social tags with professionally-created metadata. In an early study on social
tagging, Mathes (2004) first discussed the limitations and strengths of folksonomy as user-generated metadata vis-à-vis professional and author-created metadata. Bischoff et al. (2008) analyzed the overlap between social tags collected from Last.fm and music reviews written by experts for the same set of music tracks. The results showed that 73.01% of the track tags can be found inside the review pages. They also compared track tags from Last.fm to the expert created reviews from www.allmusic.com and found that 46.4% of the tags are present in the Allmusic review pages. Trant (2010) conducted a research project (steve.musum) to study the relationship of folksonomy to professionally created museum documentation. He found that 86% of user created tags was not present in museum documentation, indicating that tagging provides a significantly different vocabulary than museum documentation. He also compared tags with two controlled vocabulary sources AAT (Art and Architecture Thesaurus) and ULAN (Union List of Artists Names), which are adopted in museum’s cataloging. It turned out that many tags could not be found in both AAT and ULAN. Yi and Chan (2009) compared the social annotations collected from Delicious with Library of Congress Subject Headings (LCSH) based on word-matching. They also examined the distribution of folksonomies over the LCSH hierarchy tree. The experiment results showed that about two-thirds of social tags could be matched to LC subject headings. The study also identified three distribution patterns of tags over the LCSH tree: skewedness, multifacet and Zipfian-pattern.

Smith (2007) conducted an exploratory study to evaluate how well user-created tags matched the authoritative subject headings assigned to the same resources by comparing social tags in LibraryThing and LCSH. Based on five book samples, the author discussed the advantages and disadvantages of LibraryThing social tags and LC subject headings as
well as the efficacy and accuracy of social tags as an instrument for subject analysis. Rolla (2009) analyzed the library-assigned LCSH and LibraryThing tags for 45 books dealing with different subjects and geographic regions. He found that compared to the subject headings assigned by libraries, user-created tags can cover more aspects of a book’s subject even though the personal tags may not be effective for information retrieval. Heymann and Garcia-Molina (2009) also contrasted tags created by users on LibraryThing to the LC subject headings included in LC bibliographic records based on a large-scale dataset. Both the LibraryThing tags and the LCSH were used to annotate the same collection of books. Their study showed that LibraryThing tags and LCSH share many terms, but the usage of these terms as tags and LC subject headings was different. In this thesis, the author also compares LibraryThing tags with LCSH terms assigned by experts using the same data source but analysis is done from different perspectives. Their study analyzed the entire set of tags and LC subject headings applied to the book collection without considering their usage at book level. In the study of this thesis research, the author compares tags with LC subject headings at both collection level and book level. Besides, the author also compares tags with LCSH subdivisions and book titles.

2.2.2 Social Tags, Full Text, and Query

In order to investigate whether social tags provides an additional information source for web search or whether social tags are effective index terms, Heymann et al. (2008) analyzed the intersection of social tags with titles and textual content of web resources. Based on a sample of the Delicious data set, they found that social tags occurred in the content of half of the pages and in the titles of 16% of the pages they annotated. Bischoff
et al. (2008) did similar analysis on Delicious tags and found that 44.85% of the tags appear in the web page text. A precondition for social tags to be useful for web search is that they should intersect neatly with the query terms used for retrieving the tagged objects. Heymann et al. (2008) found a significant intersection between the queries and tags by analyzing the overlap between popular query terms from an AOL query data set and del.icio.us tags. Bishcoff et al. (2008) counted the percentage of queries containing tags on three systems: del.icio.us, Flickr and Last.fm. They found that 71.22% of del.icio.us queries consist of at least one tag and 30.61% of the queries consist entirely of tags. As for Flickr and Last.fm, the percentages are 64.54% and 12.66, and 58.43% and 6%, respectively. Rather than measuring the overlap, Suchanek et al. (2008) calculated the similarity between the frequency distributions of tags, page contents and queries based on two statistics: cosine similarity and NDGC (normalized discounted cumulative gain). The cosine similarity and NDGC were found to be 16.76% and 12.64% for tags and content and 21.61% and 25.07% for tags and queries. In this thesis, the author also compares social tags with the titles and content of tagged web documents. But the author not only studies the overlap between the two vocabularies but also studied their effectiveness as document features for clustering.

2.3 Tag-based Web Clustering

2.3.1 Clustering on Extended Document Features

Clustering algorithms have been well studied and applied in various areas in previous research (Jain et al. 1999, Leung et al. 2000, Andrews and Fox 2007). A variety of clustering approaches have been developed to discover classes and identify topics in
document collections. Traditional document clustering algorithms mostly rely on the content features contained in documents. However, as mentioned, content is not always reliable or even available, especially for web documents. Moreover, most content-based clustering algorithms, like K-means, are based on the BOW (Bag of Words) approach, which ignores the semantic relationship among words. To overcome this limitation, additional information other than document content has been used to enrich document representation during clustering process. For instance, some researchers developed heuristic similarity metrics that linearly combine link information with content information for clustering (He et al. 2001, Modha and Spangler 2000, Zhou et al. (2007). Modha and Spangler (2000) proposed an algorithm called TORIC k-means that clustered hypertext documents using words, in-links and out-links. Similarly, Zhou et al. (2007) represented each document using the combination of three vectors: in-link vector, out-link vector and text vector. Each cluster was annotated using six information nuggets: summary, breakthrough, review, keywords, citation, and reference. He et al. (2001) measured document similarity was measured based on three types of information: hyperlink structure, textual information and co-citation pattern. Instead of using link information to improve similarity metrics, Angelove and Siersdorfer (2006) developed a Relaxation Labeling based algorithm which clusters a document using its content and the cluster labels of its linked documents. A similar algorithm was also proposed by Zhang et al. (2007).

Besides links, document representation was also enriched with the background knowledge represented by ontologies. For instance, WordNet (Hotho et al. 2001) and Mesh (Yoo et al. 2006) were used as the external ontology for text enrichment. However,
they both have limited coverage. Recently, Wikipedia has been used to enhance document clustering in some studies where Wikipedia concepts and categories mapped to a document were used as additional document features (Banerjee et al. 2007, Hu et al. 2008, and Hu et al. 2009). In this thesis, we exploit how to enhance document clustering with social tags.

2.3.2 Clustering on Social Tagging

The application of social tags in clustering has been explored in several studies. Begelman et al. (2006) built a tag graph based on the co-occurrence of tags in annotated resources. A spectral bisection method was adopted to cluster the tag graph. The identified tag clusters were used for finding semantically related tags. Li et al. (2008) used association rule algorithms to identify frequent tag co-occurrence patterns, which were viewed as topics of user interests. Users and URLs were clustered under different topics based on their relation to the tags included in each topic. Although this approach can cluster both users and URLs at the same time, it has some limitations. First, the number of clusters depends on the support and the size of frequent tag co-occurrence patterns. Second, active users and popular URLs can belong to many clusters, while new or inactive users and emerging or unpopular URLs may not be assigned to any cluster. Giannakidou et al. (2008) proposes a co-clustering approach which not only utilizes tag co-occurrence but also exploits the social aspect of tagging for clustering the tagged documents. A more comprehensive research on social tagging-based clustering was conducted by Ramage et al. (2009). The authors incorporated social tags into two clustering methods: K-means and a generative clustering method based on LDA. The clustering results were also evaluated against the web directory ODP. Although their
work proved the value of tags as an additional information source for clustering, the user dimension was missing in their clustering models. Nanopoulos et al. (2009) proposed an extended spectral clustering method which clusters the tagged documents based on both the users who tagged documents and the tags they use. Tensor factorization technique was used to implement the spectral clustering on the tripartite relation among document, users and tags.

2.4 Personalized Search on Tagged Web

2.4.1 Personalized Search

In order to realize search personalization, researchers have attempted to learning user interests and preferences from a variety of sources, including web browsing and searching history (Agichtein et al. 2006, Qiu and Cho 2006, Sun et al. 2005, Tan et al. 2006, Teevan et al. 2009), users’ desktop documents and personal emails (Chirita et al. 2007, Teevan et al. 2005), as well as user explicitly specified interests (Chirita et al. 2005, Ma et al. 2007). Some of these methods have been successfully applied in industry. For instance, Google has implemented personalized search application based on users’ search and click history. However, the information sources like the users’ search, click-through and browsing history stored by specific search engine and information vendors, are normally unavailable to outsiders due to commercial and privacy reasons. Other information sources such as the emails and desktop documents are also unavailable to the public, because they are privately owned and may contain sensitive information. Comparatively, the social annotations and the structured relationships among users, social
annotations and resources provide an open and effective information channel for learning users’ information needs and interests.

2.4.2 Personalized Search based on Social Tagging

Many researchers have been interested in whether and how social annotation can be exploited for information retrieval. The potential value of social tagging for web search is empirically studied by Bishoff et al. (2008), Heymann et al. (2008), and Suchanek et al. (2008). Xu et al. (2007) developed a language model for information retrieval based on the metadata property of social tags and their relationships to the tagged documents. However, user information is not used in their model. Wu et al. (2006) proposed a probabilistic generative model in which the tags, objects and users were mapped to a common conceptual space represented with a vector. The generated conceptual vectors were used for developing various search models including personalized search. Zhou et al. (2008) proposed an extended language model based on the estimated parameters of LDA (Latent Dirichlet Allocation) for information retrieval. The proposed language model incorporated the topical background of documents and social tags as well as users’ domain interests. Harvey et al. (2011) also propose a personalized search method based on the parameters estimated from two LDA models developed on social tagging data.

Some researchers tried to implement personalized search by re-ranking the search results based on the user profiles extracted from the social tagging data. Hotho et al. (2006) proposed a page-rank like method called FolkRank for personalized search based on the social tagging graph. Bao et al. (2007) introduced two ranking methods: SocialSimRank, which ranks pages based on the semantic similarity between tags and pages, and SocialPageRank, which re-ranks returned pages based on their popularity. Xu
et al. (2008) proposed a method which re-ranks documents based on the topical similarity between the returned web pages and users who issued the query. Both web pages’ topical vectors and a users’ interest vectors were learned from the social tags associated with them. Schenkel et al. (2008) developed a top-k algorithm which re-ranks the search results based on the tags shared by the user who issued the query and the users who annotated the returned documents with the query tags. Noll and Meinel (2007) also proposed a re-ranking approach based on users’ tagging profiles. Cai and Li (2010) developed a method to build user profiles and resources profiles based on tags, which can better model users’ interests and preferences. The developed user profiles and resource profiles are used for personalized search.

Social annotations were also combined with other social data such as user blogs to construct user profiles (Bender et al. 2008, Carmel et al. 2009, Wang and Jin 2010). Bender et al. (2008) re-ranked the search results based on user topic vectors derived from user’s social tagging and blogging activities. Carmel et al. (2009) and Wang and Jin (2010) also combined the information derived from social annotation and other social systems for personalized ranking, but focused more on the social relation built through users’ social activities.

2.5 Topic Models of Tagged Web

2.5.1 Topic Analysis using Generative Models

In the data mining and information retrieval community, a set of effective probabilistic models have been proposed to simulate the generation of a document, such as the Naïve Bayesian model, the Probabilistic Latent Semantic Indexing (PLSI) model (Hofmann
1999) and the Latent Dirichlet Allocation (LDA) model (Blei et al. 2003). Particularly, the LDA model has become popular in the text mining community due to its solid theoretical foundation and promising performance. Since it was first proposed, a wide variety of its extensions have been proposed in different contexts for different modeling purposes.

Many extended LDA models have explored information other than document words for topic learning. For instance, the author-topic model proposed by Rosen-Zvi et al. (2004) used the authorship information together with the words to learn topics. The correlated LDA model learned topics simultaneously from images and caption words (Blei and Jordan 2003, Chen et al. 2009). The switchLDA model revealed topics from content words and entities in news articles (Newman et al. 2006), the Link-LDA model and Topic-Link LDA model represented topics and author communities using both content words and links between documents (Erosheva et al. 2004, Liu et al. 2009), etc. Most of these models can also be applied in the social annotation context when considering the tags as the additional information source for topic learning.

2.5.2 Topic Models of Tagged Web

A variety of LDA-based generative models have been proposed for modeling the generation of social tags. For clustering purpose, Ramage et al. (2009) proposed a LDA model which jointly models the generation of content word and tags. This model is essentially the same as the Conditionally-independent LDA (CI-LDA) model used for generating words and entities by Newman et al. (2006) and the Link LDA model used for generating words and document links by Erosheva (2004). In CI-LDA model, the tag was generated from the same source as the word: the topics of the document. Users’
impact on the generation of tags was not considered in this model.

Zhou et al. (2008) proposed a more comprehensive model for social annotation, which considered the impact of both document topic and user interest on tag generation. In this model, first, a user decided to annotate a web document and then the user selected a topic, based on which a tag was generated to describe the document. Although this model provides a comprehensive view about the generation process of both content words and tags, unfortunately, it is intractable due to too many parameters to be estimated. Therefore, Zhou et al. simplified the model by assuming that words and tags are both generated from the same topics shared by documents and users, which is not proper because document words are created by the authors of the document, while tags are collaboratively generated by users with different background and interests.

Kashoob et al. (2009) proposed a generative model called community-based categorical annotation (CCA) model. Different from other models, in this model, the annotation process was modeled as a collective decision of user communities, which were viewed as groups forming around interests, expertise, language, etc. It was assumed that each community has a number of underlying categories as its world view, and each category could be represented as a mixture of tags. Therefore, in CCA model, a tag in a document was generated from a category which was further generated from a community selected for the document. The outputs of this model included the community distribution of a resource, category distribution of a community, and tag distribution of a category. The authors compared the category distribution generated from the CCA model and the hidden topics in content words generated separately through standard LDA model. They concluded that tag-based categories were not the same as content-based topics. A defect
of this model is that it ignored the dependence of tags on resource topics. According to this model, tags are generated independently from resource topics. Apparently, this is not the case in real tag generation process. Moreover, this model only considered the collective impact of communities on social annotation. However, in some cases like personalized search, we are more interested in the information about individual users.

Bundschus et al. (2009) adapted the correlated or correspondence LDA (CorrLDA) model proposed by Blei and Jordan (2003) for modeling social tagging data. The CorrLDA model first generated word topics for a document. Then the topics associated with the words in the document were used to generate tags. Compared to the CI-LDA model, the CorrLDA model can force a greater degree of correspondence between two information sources (in this case, words and tags). But like CI-LDA, the user information is missed in the tag generation process. In order to incorporate user factors into the tag generation process, another model called User-Topic-Tag Model was proposed (Bundschus et al. 2009). In this model, users were treated like authors in the author-topic model proposed by Rosen-Zvi et al. (2004). First, for each word in the document, a user was chosen uniformly at random from the group of users who annotated the documents. Then, a topic was chosen from a distribution over topics specific to that user, and the word was generated from the chosen topic. Finally, as in the CorrLDA model, the topics associated with the words in the document were used to generate each tag associated with the document. Although this model accounted for user factors, it did not correctly simulate the real social annotation process because users were modeled as creators of content words instead of tags.
CHAPTER 3: A COMPARISON BETWEEN SOCIAL TAGS AND EXPERT-ASSIGNED SUBJECT TERMS

3.1 Introduction

Social tagging can be viewed as “a process by which many users add metadata in the form of keywords to shared content” (Golder and Huberman 2006). In other words, social tagging provides an open and collaborative approach for metadata creation. Nevertheless, user-created tags have been criticized for imprecision and semantic ambiguity caused by the prevalence of non-words, polysemy and synonymy, as well as a lack of hierarchy (Mathes 2004, Golder and Huberman 2006, Suchanek 2008). Macgregor and McCulloch (2006) point out that these quality problems involving social annotations originate from the absence of properties that characterize controlled vocabularies.

Traditionally, metadata is created by information professionals such as cataloguers based on controlled vocabularies. For instance, when applying a subject term to a resource, a cataloguer first summarizes the subject content of the resource and then matches the subject content of the resource with the best subject representation captured in a controlled vocabulary such as the Library of Congress Subject Headings (LCSH). Expert created metadata is considered of high quality. It plays an indispensable role in libraries and other information institutions. It is still the foundation of digital libraries and museums. Therefore, it is proper to use expert created metadata as the benchmark to check the quality of social tags and their effectiveness as index terms.

Despite of the claimed high quality, expert created metadata nevertheless also has several limitations. First, it is costly to produce and difficult to scale. Second, the
language used by information professionals such as subject headings is often highly
technical and specialized. Users who want to get access to the resources by searching the
catalogue by subject are expected to be knowledgeable about the specialized subject
headings. This stringent requirement renders many professionally indexed resources
inaccessible to regular users (Trant, 2010). On the other hand, the great success and
popularity of modern search engines to a great extent rely on the fact that users can obtain
desired results by entering free-form keywords. Moreover, an early study of the
vocabulary problem in human–system communication has shown that ‘no single access
word, however well chosen, can be expected to cover more than a small proportion of
users’ attempts’ (Furnas et al. 1987). Based on several experiments, Furnas et al. (1987)
showed that there were no rules, guidelines or procedures for choosing a good name, in
the sense of “accessible to the unfamiliar user”. They further pointed out that ‘the only
hope for untutored vocabulary driven access is to provide many, many alternate entry
terms’. Based on these conclusions, it seems that social tagging provides an opportunity
for information institutions to expand the access points of their resources beyond
professionally created index terms and thus improve the accessibility of their resources.

In social tagging system, the entrance threshold of metadata creation is lowered and
a great number of resources can be tagged within a short time. Besides, unlike
professionals who create metadata, taggers annotate resources from multiple individual
perspectives. Resource organization is not the only motivation of users’ tagging behavior.
As pointed out by Zollers (2007), taggers are also motivated by social incentives. Zollers
(2007) identified six incentives of social tagging: future retrieval, contribution and
sharing, attracting attention, play and competition, self-presentation and opinion
expression. Users are always motivated by a number of these incentives simultaneously during the tagging process. To some extent, the user-created tags provide a window into users’ interests, behaviors and attitudes that might help information institutions better understand and serve users.

Despite the quality issues of social annotation caused by its uncontrolled nature, many library researchers have realized the unique value of social annotation and propose the co-existence of social tags with controlled vocabularies like LCSH. Macgregor and McCulloch (2006) stated that “librarians and information professionals have lessons to learn from the interactive and social aspects exemplified by collaborative tagging systems as well as their success in engaging users with information management”. Rolla (2009) pointed out that tags can be used to enhance subject access to library collections.

In this chapter, the author conducts a study to investigate the differences and connections between social tags and expert-created subject terms, which are an integral part of library metadata. This study is designed to shed lights on questions like:

- Whether social tags are comprised of a vocabulary similar to expert-assigned subject terms?
- What are the limitations of social tagging compared to the professional cataloging approach?
- In what aspects can social tags and expert-created subject headings complement each other?

The rest of the chapter is organized as follows: Section 3.2 introduces the datasets used for analysis. Section 3.3 describes the data analysis process and the results obtained. Section 3.4 summarizes the conclusions and gives a brief overview of future work.
3.2 Dataset

The dataset was collected using the same approach as in (Heymann and Garcia-Molina 2009). First, a dump of Library of Congress bibliographic records in MARC format was downloaded from the Internet Archive\(^7\). These MARC records are encoded in XML format. We randomly selected 114,598 records for experimentation. Each bibliographic record comprises many MARC fields. Figure 3.1 shows an example of the bibliographic record (some data fields are omitted). In our study, for each record the author only extracted the content of Field 020 (ISBN), Field 245 (Title) and Field 650 (Topical Heading). The records without these fields were filtered out. Field 650 is not the only subject access field which can occur in a MARC record. There are many other fields (6XX) which can hold subject information. However, in the dataset, most MARC records only use Field 650 to represent the subject information as shown in Figure 3.1.

The subject terms included in Field 650 are assigned by expert cataloguers based on a generally accepted subject heading system such as the LCSH, Medical Subject Headings and Canadian Subject Headings. In our dataset, the subject entries of Field 650 in all bibliographic records conform to LCSH. LCSH has subject headings and subdivisions. When encoded in MARC, the LCSH subject headings and subdivisions are structured into several subfields of 650. The terms contained in several subfields together form a full subject heading. For instance, the book record shown in Figure 1 has three 650 fields representing three subject headings which are composed of the terms contained in their subfields. In this study, the author treat the terms contained in each subfield as separate keywords. Tags are compared with these separate subject terms.

\(^7\) http://www.archive.org/
Figure 3.1  An example of a bibliographic MARC record

The ISBNs of the selected records were used to crawl tags assigned by users to corresponding books on the social tagging website LibraryThing. The crawling process was conducted during October 2009. In total, 39,576 books were found to be annotated on LibraryThing. Many books were annotated with a few of the tags. In order to reduce the sparseness of the dataset, the author filtered out those books with fewer than 20
unique tags. The tags applied to the books with at least 20 unique tags can cover more than 80% of the tags applied to all the books. The resulting dataset contains 8562 book records used for experiment and analysis.

In total, the 8562 book records were annotated with 176,105 unique tags by users and 7628 unique Library of Congress subject terms by experts. We can see that the size of the tag vocabulary is much larger than the LCSH vocabulary. This is because tags are created by a large number of users with different backgrounds in a totally uncontrolled environment. Different users may use different terms to describe the same subject. LCSH terms are chosen by information professionals from a controlled vocabulary, which is a carefully selected list of words and phrases. In addition, the number of subject headings assigned to each book is limited whereas the book can be annotated any number of times and with any number of different tags in the social tagging system. In most cases, fewer than five subject headings are assigned to a work. Therefore, the number of unique terms included in the subject headings assigned to each book is always much smaller than the number of tags assigned to the books.

3.3 Analysis and results

For analysis, social tags and LCSH terms are compared from different aspects. Generally, the comparison is conducted at two levels: the collection level and book level. At the collection level, the author compares the social tags and the subject terms present in the whole dataset without considering their application to specific books. At the book level, social tags are compared to subject terms applied to the same books. In the word-comparison process, the author only considers the syntactic form of the terms. In other words, a social tag and a LCSH term are considered equivalent only if they are syn-
tactically identical. For instance, the tag ‘historical fiction’ and the LCSH term ‘Historical fiction’ are considered equivalent. But the tag ‘novel’ and the LCSH term ‘fiction’ are viewed as two different terms even though they share the core semantic meaning.

3.3.1 Overlapping vocabulary

The author first compares the entire set of unique tags and LCSH terms applied to the entire collection of books. It is found that 3824 unique terms have not only been adopted as tags by users but they are also used in LCSHs by the experts. These overlapping terms only comprise a small portion (about 2.2%) of the entire tag vocabulary. That is, about 97.8% of tags cannot be found in LCSHs for the same collection of books. However, the overlapping terms compose about half (50.1%) of the terms applied in LCSHs. This indicates that, given a LCSH, there is a 50% probability that it has also been adopted by users as a social tag.

The author checked the frequency or popularity of the overlapping terms in tags and LCSHs. The frequency or popularity of a term is measured by the number of books annotated by it. The more popular a tag, the more frequently it is used to annotate an array of books. Table 3.1 lists the average frequency of the shared terms and the average frequency of the entire set of social tags and LCSH terms. From the table we can see that the average frequency of the overlapping terms is higher than the average frequency of the whole vocabulary, especially for tags. The average frequency of overlapping terms as tags is 33.5, but the average frequency of the whole set of tags is 5.3. This means that the tags which can be found in LCSHs are on average used more often by users.
Table 3.1  Average frequency of shared terms and the whole set of tags and LCSH terms

<table>
<thead>
<tr>
<th></th>
<th>Overlapping Terms</th>
<th>All Terms</th>
</tr>
</thead>
<tbody>
<tr>
<td>Social Tag</td>
<td>33.5</td>
<td>5.3</td>
</tr>
<tr>
<td>LCSH</td>
<td>7.0</td>
<td>4.6</td>
</tr>
</tbody>
</table>

The author further checked the popularity rank of the overlapped terms in social tags and LCSHs. Specifically, the author wants to know whether a shared term is equally popular when used as tags and subject terms. The author ordered the overlapping terms according to their frequency in tag vocabulary and LCSHs respectively, and computed the rank correlation of these two rank sequences. The Spearman’s rank correlation coefficient was adopted, which assesses how well the relation between two rankings \(X\) and \(Y\) can be described using a monotonic function. Its value is inside the interval \([-1, 1]\). If \(Y\) tends to increase when \(X\) increases, the Spearman correlation coefficient is positive. If \(Y\) tends to decrease when \(X\) increases, the Spearman correlation coefficient is negative. A greater absolute value of Spearman correlation coefficient indicates a stronger correlation. A Spearman correlation of zero indicates that the two rankings are completely independent. In our case, the Spearman correlation coefficient of the two rankings (the ranked overlapping terms according to their frequency in tags and LC subject headings) is \(0.507\). This indicates a positive but not very strong correlation between the two rankings. Namely, if an overlapping term is frequently used by users as a tag, there is about a 50% probability that it is used equally frequently in LCSH by experts.
3.3.2 Frequent tags and LC subject headings

To further examine whether tags share the professional vocabulary of subject experts or represent the different vernacular vocabulary of users, in this section the author compares top \( n \) frequent tags with top \( n \) frequent LCSH terms. The Jaccard similarity coefficient (also called Jaccard index) is used to measure the similarity between the frequent sets of tags and LCSH terms and is calculated according to Equation 1:

$$J^{(n)}(T, L) = \frac{|T \cap L|}{|T \cup L|}$$  \hspace{1cm} (3.1)

In equation 3.1, \( n \) is the number of top frequent terms. \( T \) is the set of top \( n \) frequent tags and \( L \) is the set of top \( n \) frequent LCSH terms. Figure 3.2 shows the Jaccard index when \( n \) varies from 100 to 1000. We can see that the Jaccard index is always between 0.08 to 0.1, indicating a very low overlap between the top frequent tags and top frequent LCSH terms. This means that the frequent terms used by users and experts are very different. Even though some terms are very popular among users, they are not frequently used by experts, and vice versa.

![Jaccard Index for Top frequent tags and LCSH terms](image-url)
Table 3.2  Top 30 frequent tags and LC subject headings

<table>
<thead>
<tr>
<th>Tags</th>
<th>LCSHs</th>
</tr>
</thead>
</table>

Table 3.2 lists the top 30 popular tags and Library of Congress subject terms. There are only three terms, ‘fiction’, ‘history’ and ‘20th century’, which are used as both tags and Library of Congress subject terms. The LCSH terms contain many phrases, such as ‘juvenile literature’ and ‘history and criticism’ while tags usually consist of one word. Moreover, all LCSH terms are used to describe the subjects of the works whereas many tags are used for purposes other than subject description. Some tags are used for self-reference, such as ‘read’, ‘unread’, ‘to read’, ‘read in 2009’ and ‘owned’. Some tags are used to describe physical features, such as ‘hardcover’ and ‘paperback’. Also, there are tags about edition information, such as ‘first edition’. The tags like those for self-reference may be useful for users who created them but do not fit the traditional cataloguing and classification systems. Previous studies of Delicious also found that while a number of tags are subject related and ‘good enough’ for index terms, many others are not directly related to the subject (Golder and Huberman 2006, Kipp and Campbell 2006). In a study, Kipp and Campbell (2006) found that over 16% of Delicious
tags were non-subject terms. In another study, Kipp (2007) examined the use of two categories of non-subject tags: affective tags and time- and task-related tags. She suggests that the use of non-subject tags shows that users wish to attach personal resource management information to documents and view classification as a holistic process closely tied to themselves and their work.

### 3.3.3 Tags and LC subject headings annotating the same book

In the above analysis, tags and LCSH terms are compared at the collection level. In this section, the author analyzes tags and LCSH terms at the book level. Tags applied to an individual book are compared to the LCSH terms assigned to the same book. Overall, there are 7276 (85%) book records whose LCSHs contain at least one term which is also used by the users to tag the same books. This means that in most cases users and experts agree on at least one term which may be used to describe the subject of a book.

Table 3.3 Number of book records with different percentages of LCSHs present in tags

<table>
<thead>
<tr>
<th>Percentage of LCSHs overlapped by the tags for each book</th>
<th>≥80%</th>
<th>≥70%</th>
<th>≥60%</th>
<th>≥50%</th>
<th>≥40%</th>
<th>≥30%</th>
<th>≥20%</th>
<th>≥10%</th>
<th>&lt;10%</th>
<th>total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of books</td>
<td>3</td>
<td>29</td>
<td>102</td>
<td>407</td>
<td>987</td>
<td>2088</td>
<td>4329</td>
<td>7030</td>
<td>1532</td>
<td>8562</td>
</tr>
<tr>
<td>Percentage of books</td>
<td>0.04%</td>
<td>0.34%</td>
<td>1.19%</td>
<td>4.75%</td>
<td>11.53%</td>
<td>24.39%</td>
<td>50.56%</td>
<td>82.11%</td>
<td>17.89%</td>
<td>100%</td>
</tr>
</tbody>
</table>

The author calculated the number of books with different percentages of LCSH terms covered by user tags (see Table 3.3). As shown in Table 3.3, for more than 80% of the books, at least 10% of each book's LCSH terms can be found in its tags; for more than
half of the book records, at least 20% of each book's LCSH terms are covered by its tags. For 407 books, more than half of each book's LCSH terms are also used as its tags.

Table 3.4  Example terms assigned by users and experts to the same books and different books

<table>
<thead>
<tr>
<th>Terms usually used to annotate the same books by users and experts</th>
<th>Terms used by users and experts to annotate different books</th>
</tr>
</thead>
<tbody>
<tr>
<td>fiction, biography, history, united states,</td>
<td>adult, office, women’s studies, theory, mystery</td>
</tr>
<tr>
<td>England, women, 20th century, Christianity,</td>
<td>fiction, occult, paper, domestic fiction, drugs</td>
</tr>
<tr>
<td>19th century, juvenile fiction, friendship,</td>
<td>young adults, translations, plays, diet</td>
</tr>
<tr>
<td>English language, France, magic, philosophy,</td>
<td>paganism, Romans, texts, government policy,</td>
</tr>
<tr>
<td>presidents, police, murder, dogs</td>
<td>computer software, methods</td>
</tr>
</tbody>
</table>

The author calculated the number of unique terms which have been used by both users and experts to describe the same book at least once. In total, there are 3083 such terms. When users and experts use a term to annotate the same book, it indicates that they have some common understanding about the semantics and usage of the term. The more times a term is used by users and experts relating to the same books, the more likely it is that users and experts agree on the meaning and usage of the term. The left column of Table 3.4 lists some terms which are often applied by users and experts to the same books in our dataset. However, users and experts may also apply the same terms to different books. As shown in Section 3.3.1, 3824 terms are used both by users as tags and subject terms by experts. However, only 3083 of them are used by taggers and experts to annotate the same books. This means that 741 (3824–3083) terms are used in different ways by users and experts. An example is ‘mystery fiction’, which appears in both tag vocabulary and LCSHs. However, users assign this term to many books (such as ‘Jack
and Jill’, ‘Cat and Mouse: a Novel’, ‘River of Darkness’) which are not annotated by experts with this term. On the other hand, the only two books (‘The 8th Confession’ and ‘The Problem Child’) whose LCSH contains this term are not tagged with this term by users. Other example terms like ‘mystery fiction’ are listed in the right-hand column of Table 3.4.

### 3.3.4 Tags and LCSH subdivisions

Each 650 data field comprises several subfields. Besides the main topical subject, in MARC coding, field 650 also uses subfield to parse LCSH subdivisions. The LCSH subdivisions are used to narrow the scope of a heading and bring out specific aspects of a subject. In the above analysis, social tags are compared with the terms contained in all 650 subfields. In this section, the author compares social tags with terms contained in each 650 subfield separately. In our dataset, the 650 field generally contains the following subfields:

- $a$ - topical or geographic name entry element
- $x$ - topical subdivision
- $y$ - chronological subdivision
- $z$ - geographic subdivision
- $v$ - form subdivision

The first subfield ($a$) designates a topical subject or a geographic name as an entry element. The other four subfields are subdivisions: form subdivision ($v$) designates a specific kind or genre of material; topical subdivision ($x$) usually represents a particular subtopic and narrows down the broader subject; chronological subdivision ($y$) represents a period of time; and geographic subdivision ($z$) specifies geographic information.
Table 3.5  Statistics of LCSH subfield terms

<table>
<thead>
<tr>
<th>LCSH subfields</th>
<th>$a$</th>
<th>$v$</th>
<th>$x$</th>
<th>$y$</th>
<th>$z$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total Number of terms</td>
<td>6615</td>
<td>123</td>
<td>634</td>
<td>66</td>
<td>449</td>
</tr>
<tr>
<td>Number of terms used as tags</td>
<td>2639</td>
<td>42</td>
<td>236</td>
<td>12</td>
<td>281</td>
</tr>
<tr>
<td>Percentage of terms used as tags (%)</td>
<td>39.9</td>
<td>34.1</td>
<td>37.2</td>
<td>18.2</td>
<td>62.6</td>
</tr>
</tbody>
</table>

Table 3.6  Top 10 LCSH subfield terms also used as tags

<table>
<thead>
<tr>
<th>$a$</th>
<th>$v$</th>
<th>$x$</th>
<th>$y$</th>
<th>$z$</th>
</tr>
</thead>
<tbody>
<tr>
<td>women</td>
<td>fiction</td>
<td>History</td>
<td>20th century</td>
<td>united states</td>
</tr>
<tr>
<td>friendship</td>
<td>biography</td>
<td>christianity</td>
<td>19th century</td>
<td>england</td>
</tr>
<tr>
<td>fantasy</td>
<td>juvenile fiction</td>
<td>philosophy</td>
<td>18th century</td>
<td>france</td>
</tr>
<tr>
<td>English language</td>
<td>drama</td>
<td>grammar</td>
<td>21st century</td>
<td>new york</td>
</tr>
<tr>
<td>magic</td>
<td>patterns</td>
<td>rhetoric</td>
<td>16th century</td>
<td>great britain</td>
</tr>
<tr>
<td>presidents</td>
<td>poetry</td>
<td>poetry</td>
<td>17th century</td>
<td>germany</td>
</tr>
<tr>
<td>police</td>
<td>humor</td>
<td>usage</td>
<td>2008</td>
<td>italy</td>
</tr>
<tr>
<td>murder</td>
<td>interviews</td>
<td>drama</td>
<td>1933-1945</td>
<td>india</td>
</tr>
<tr>
<td>dogs</td>
<td>juvenile literature</td>
<td>English</td>
<td>15th century</td>
<td>europe</td>
</tr>
<tr>
<td>orphans</td>
<td>early works to 1800</td>
<td>judaism</td>
<td>1800</td>
<td>california</td>
</tr>
</tbody>
</table>

The author compared social tags with terms contained in each of these five LCSH subfields to investigate whether users are more likely to annotate books with terms belonging to a certain subfield. The first row of Table 3.5 lists the total number of unique terms present in each LCSH subfield in our dataset. The aggregated number of terms contained in these subfields is larger than the total number of unique subject terms 7628. This is because some terms appear in more than one subfield. For instance, the term ‘fiction’ is found in both the topical subdivision ($x$) and the form subdivision ($v$). The
second and third rows of Table 3.5 respectively show the number and percentage of LCSH subfield terms which are also applied by the users to annotate the same books. We can see that the chronological subdivision ($y$) has the lowest percentage of terms used as social tags whereas the geographic subdivision ($z$) has the highest percentage of terms which are adopted by users as tags. More than half of the terms appearing in the geographic subdivision are adopted by users as social tags to annotate the same books. This indicates that users are most likely to agree with experts on terms designating geographical information. The low percentage of chronological terms present in tags can be caused by either of two reasons: users do not tag books from the chronological aspect, or users and experts use different chronological terms to annotate the books. Whatever the reason is, since LCSH contains many chronological terms which are not covered by tags, it can be used to complement the tag-based system when users retrieve books according to time period. Table 3.6 lists the top 10 subfield terms which are also applied by users to annotate the same books.

For each subfield, we calculated the number of catalogue records which have at least one term present in social tags. The results are shown in Table 3.7. For instance, among the catalogue records containing the subfield $a$ (topical term or geographic name entry element), 62.1% have at least one term in $a$ adopted by users as a tag to annotate the books. Except for $y$ (chronological subdivision), all subfields at least partly overlap with social tags in more than half of the book records. The percentage of book records with the chronological subfield covered by tags is still lowest, under 50%. This further indicates that users are less likely to annotate books with chronological terms or agree with the experts on terms representing the chronological information.
Table 3.7 Number of book records which have at least one tag appearing in LCSH subfields

<table>
<thead>
<tr>
<th>LCSH subfields</th>
<th>$a$</th>
<th>$v$</th>
<th>$x$</th>
<th>$y$</th>
<th>$z$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of records with this field</td>
<td>8562</td>
<td>4052</td>
<td>3325</td>
<td>678</td>
<td>2590</td>
</tr>
<tr>
<td>Number (%) of records which have at least one tag appearing in this field</td>
<td>5371 (62.7)</td>
<td>3156 (77.9)</td>
<td>1870 (56.2)</td>
<td>315 (46.5)</td>
<td>1527 (59.0)</td>
</tr>
</tbody>
</table>

3.3.5 Tags, LC subject headings and book titles

Book title is a major access point in all catalogues. For users who have a specific title in mind, the easiest and most effective way to find the book is to search by title. But in most cases, users with an interest in a topic do not know the specific titles of the books under a given topic. Therefore, topic-based search is essential. Because book titles already consist of terms which the authors think most appropriate to reflect the subject of their books, to be able to enhance the subject access, social tags and LC subject headings should contain terms beyond titles. In this section, the author compares tags and LCSH terms with the titles of annotated books to see whether users and experts are able to provide terms beyond those already included in the titles.

The author calculated the tags and LCSH terms which occur in the titles of the books annotated by them. Results indicate that 4119 (2.3%) unique tags and 812 (10.6%) unique LCSH terms can be found in the titles of the books annotated by them. The author also calculated the number of books which have at least one tag or LCSH term appearing in their titles. It turns out that 5149 (60.1%) books have at least one tag correspondent to
a term in their titles, and 1529 (17.9%) books have at least one LCSH term present in their titles. It is clear that, although only a tiny portion of tags appear in book titles, more than half of the books have at least one tag coming from the title. On the contrary, despite the fact that the percentage of LCSH terms overlapping with titles is higher, only a minority of books have LCSH terms appearing in the book titles. This is because subject experts intentionally avoid repeating the title in subject terms. Users are more likely than experts to annotate books with terms from the titles. However, users also use a much greater variety of terms beyond titles to annotate the books. Therefore, overall user-created tags and expert-generated LCSH can both complement the title-based search. Users enrich the access points by annotating books with a multitude of different terms whereas experts enhance the subject access by annotating books with subject terms not covered by book titles.

3.4 Conclusion

In this chapter, the author investigated how social tags are different from the subject terms created by professional cataloguers. The study sheds light on the value, feasibility and obstacles of using social tags as index terms. The results of this study also provide some tangible evidence for those who are concerned with the feasibility and effectiveness of social tagging as an indexing approach. The conclusions drawn from the results are summarized as follows.

Users and experts share some common terms for annotating books. About half of the LCSH terms are covered by social tags. In most cases, users and experts agree on at least one term. However, tags mostly comprise a vocabulary different from that of professionals. Only a tiny portion (2.2%) of tag vocabulary overlaps with LCSH terms.
Even overlapping terms may be used by users and experts in different ways. A term frequently used by users may only be used by experts a few times, and vice versa. Additionally, the same term can be used by users and experts to annotate different books. Moreover, analysis of popular tags and subject terms show that the top frequent terms used by users and experts are very different. These findings show that social taggers might help enhance the subject access to library collections by describing library resources with terms different than those used by experts, although further study may be needed to determine whether and to what extent the tags not found in subject terms are useful.

An identified obstacle for social tagging to be formal indexing approach is the prevalence of tags which are used for purposes other than topical description. Subjective and personal tags, such as ‘to read’, ‘read in 2009’ and ‘unread’ are among the top frequent tags. These tags show that users view classification as a holistic process closely tied to themselves and their work (Kipp 2007). These tags may be useful for the users who created them but they do not fit into the traditional cataloguing and classification systems.

LCSHs bring together the topical, geographical, chronological and genre aspects of a work. Each aspect is represented with a subfield. In order to ascertain how these subfields are represented in social tags, the author compared social tags with the terms included in each subfield. The results show that the chronological subdivision of LCSH is least covered by social tags. This indicates that users mostly do not annotate books from a chronological aspect or they tend to use chronological terms differently from those in LCSH to represent a period of time. If users always use different terms to describe the
chronological information of books, then social tags and expert-created subject terms can complement each other to expand the chronological access points of book collections.

Book titles include terms which the authors consider most representative of the subject matter of the work. According to data analysis of the terms in titles which are used as tags and LCSH, users are more likely than experts to annotate books with terms from the titles. More than half of the books have at least one tag present in the title in our dataset. However, users also annotate books with terms other than those contained in book titles. In that sense, both user-created tags and expert-generated subject headings can greatly complement the title-based search.

A major limitation of the study derives from the fact that comparison analysis is only at the syntactical level. Terms which have different syntactic forms but similar semantic meanings are prevalent. Taking into account the semantics of the tags may lead to different findings. For a future work, the author can conduct a comparison analysis of tags and LCSH on the semantic level to ascertain if users and experts apply the terms with more or less the same meaning to different books and to identify the semantic differences between the tags and LCSH applied to the same book. In addition, it will be useful to investigate whether user tags overlap with the terms used in search queries, especially those tags absent from expert-assigned subject terms.

3.5 Research Question Tested

Question 1: Are social tags effective document features which can be used to represent and index web documents in web mining and search applications?

In this chapter, the author used expert created subject headings as the benchmark to investigate the effectiveness of social tags as index terms.
CHAPTER 4: A COMPARISON OF SOCIAL TAGS, AUTHOR-PROVIDED METADATA AND CONTENT WORDS

4.1 Introduction

Before social tagging, author-created metadata were seen as an appropriate alternative of professional-generated metadata for web documents (Greenberg et al. 2001, Mathes 2004). Authors can add metadata to their work in many forms. Common examples of author-created metadata include keywords, subject descriptors, categories, descriptions, and abstracts of web documents. Compared to professional-created metadata, author-generated metadata is highly scalable and less costly. Besides, authors are most intimate with their work and thus supposed to be able to create effective metadata to describe their work (Greenberg et al. 2001). Moreover, authors of web documents are motivated metadata creators because they would want their work to be discovered and referred by others on the web. If they understand the metadata would increase the availability and publicity of their work, the authors would be willing to provide relevant metadata for their work.

Despite of the merits of authors as metadata creators, some researchers point out that authors are not necessarily any better at generating metadata than anyone else (Mathes 2004, Thomas and Griffin 1999). Like social tags, author-generated metadata has also been criticized for poor quality. Authors were considered to be lacking in expert skills and incapable of producing high quality metadata by some researchers (Weinheimer 1999). Besides, Mathes (2004) pointed out that “author created metadata may help with
the scalability problems in comparison to professional metadata, but both approaches share a basic problem: the intended and unintended eventual users of the information are disconnected from the process”. In other words, although authors are most familiar with their work, they may have different understandings from the users who use their work, and the metadata or index terms they provide for their work may not be the same terms used by people to search their work. In social tagging, this problem seems to be solved, because taggers are at the same time indexers and searchers. Therefore, we could expect that the tags applied to a web document are very likely to be used as the search query to retrieve the same web document.

In order to further examine the quality of social tags as index terms and find out whether users are better metadata creators than authors, in this chapter, the author compares social tags assigned to web pages with the keyword and description metadata provided by web page creators. Both of them are also compared with the content of web documents. This is because most current web mining and search applications are based on full-text indexing. It is useful to see whether social tags and author-provided metadata can be utilized to improve content-based applications. The three sources of index terms (social tags, author-provided metadata and web content) are compared both at the vocabulary level using word-match method and at the application level with clustering methods.

The comparison is also conducted to examine the metadata quality of social tags and author-provided keywords and description. Guy et al. (2004) state that “high quality metadata supports the functional requirements of the system it is designed to support, which can be summarized as quality is about fitness for purpose”. In this chapter, the
author focuses on two major functions of metadata defined by the National Information Standards Organization (2004): (1) providing additional information of the resource it describes; (2) facilitating the discovery of relevant information by bringing similar resources together and distinguishing dissimilar resources. Specifically, in this chapter, high-quality metadata refers to the metadata which provides additional information about the topic of web pages and group topically similar web pages into clusters.

In the study, real-word data were collected for comparison analysis. The dataset for social tags were collected from the Delicious social bookmarking system. The dataset for author-provided metadata consists of the <meta> attributes (“keywords” and “description”) embedded in the HTML or XHTML documents of the web pages annotated by Delicious tags. Social tags and author-provided keywords and description terms are compared in different aspects.

The rest of the chapter is organized as follows. Section 4.2 introduces the details of the dataset. Section 4.3 investigates whether user tags and author-provided keywords and descriptions provide extra information beyond web content on word match analysis. Section 4.4 assesses the effectiveness of user-created tags and user-generated metadata for web clustering. The K-means clustering method is used to explore whether clustering performance can be enhanced by incorporating tags or author-provided keywords and descriptions into the clustering process. The clustering results are evaluated against a user-maintained web directory (Open Directory Project—ODP) based on several quality metrics.
4.2 Dataset

The Delicious dataset introduced in section 1.4 is used to collect the social tags for analysis. The tagged URLs are used to extract the web content and author provided metadata. Specifically, for the tagged URLs, the author crawled the value of “keywords” and “description” attributes provided within the <meta> tag in the head section of their HTML or XHTML documents. The <meta> elements are used by web page authors to provide structured metadata about their pages. The two most common attributes of meta elements are “description” and “keywords”. For the same web pages, the author also extracted their title and the textual content for comparison.

In the comparison analysis, clustering method is adopted to examine the value of social tags, as well as author provided keywords and description. However, there is no golden standard for clustering evaluation. To solve this problem, following the approach adopted in (Ramage et al. 2009), the author used the web categories of the Open Directory Project (ODP)* as the ground truth of web clusters. ODP is a human-edited hierarchical Web directory containing 17 top-level categories. In the study, only the 14 top-level categories were selected as the clustering standard (see Figure 4.1). By overlapping the URLs from the Delicious dataset with the URLs listed under the 14 categories of ODP, 45,462 URLs tagged in 208,437 Delicious bookmarks were obtained. It was found that 23,188 of the 45,462 URLs contained both “keywords” and “description” attributes. Figure 4.1 shows the distribution of the 23,188 URLs among the 14 ODP categories. Note that a URL may belong to more than one category.

* http://www.dmoz.org/
In the data set, the 23,188 URLs are associated with 19,934 different tag terms and 90,827 different keyword terms. We can see that the number of different keyword terms is about five times larger than the number of different tags. This owes to the fact that the keywords were generated by the authors of each URL while the tags were collaboratively created by users. In order to reveal the usage patterns of the keywords and tags applied to the URLs, the author analyzed their frequency distribution. Figure 4.2 and Figure 4.3 respectively show the distributions of tags and keywords with different frequency. Since both axes of the figures are in log scale, Power Law distributions are clearly seen. The power law distribution indicates that a relatively small number of tags are highly used by users while most tags are only applied to a few objects. Seemingly, a very small portion of keyword terms are frequently used by authors to describe the web pages they create while most keyword terms are only used to describe a limited number of web pages.

Figure 4.1 Distribution of the 23,188 URLs over 14 ODP categories
4.3 Vocabulary Analysis

In this section, the author explores whether social tags and author-provided keywords and descriptions add new information to existing web page content by vocabulary analysis. Presently, most web search engines primarily rely on page content and link structure for indexing, searching and ranking. For social tags and author-provided metadata to have impact on improving the performance of web search or other web mining tasks, they
should contain useful information which cannot be extracted from the page content directly. To investigate whether social tags and author-provided keywords contain additional information value, we examine their occurrence in the title and content of the annotated web pages based on word match.

- **Tag vs. Title.** Among the 19,934 different tags assigned to the 23,188 URLs, 4854 (24%) of them appear in the titles of the annotated pages. 12,482 (54%) URLs are tagged by at least one term which can be found in their titles; 2024 URLs have half or more tags occurring in their titles.

- **Keywords vs. Title.** Among the 90,827 different keyword terms, 14,965 (16%) can be found in the title of the web pages they describe. In addition, 19,857 (85.6%) URLs have at least one keyword term present in the page title; 3344 (14%) URLs have at least half of the keyword terms present in their titles.

- **Tag vs. Content.** Among all tags, 8,430 (42%) can be found in the textual content of the pages annotated by them. About 80% (18,568) of the annotated pages have at least one tag present in their page text.

- **Keywords vs. Content.** 42,438 (47%) of the keyword terms are present in the textual content of the pages they describe. In addition, for 22,178 (96%) of the URLs, at least one of the keyword terms is present in the page text.

Based on the analysis, we can see that less than half of the tags or keywords are present in the title or content of the pages. This indicates that both tags and keywords contain additional information value beyond the title and page content of the pages. However, tags are more likely than keywords to be found in web page titles. This may be because some authors deliberately avoid choosing terms already used for page titles as
keywords.

However, for about 85% of the URLs, keywords provided by authors contain at least one term present in the title, whereas only about half of the URLs are annotated by at least one tag present in the page title. Authors are also more likely than users to adopt terms from the page content to annotate the pages. Almost all (96%) URLs are described by at least one keyword term from the page content.

The analysis above reveals some difference in the usage patterns of tags and author-provided keywords. In order to further examine whether users and authors tend to use different terms to describe the same page, the rest of this section examines the overlap between user tags and author-provided keywords and descriptions for the same pages.

- **Tag vs. Keywords.** 6,447 terms (32% of tags, 7% of keywords) are applied as both tags by users and keywords by authors to describe the same page. 16,113 (70%) URLs have at least one tag that is also adopted by the authors as the keyword to describe the pages.

- **Tag vs. Description.** 5,634 (28%) tags appear in the descriptions provided by the authors for the same pages. 15,161 (65%) URLs are annotated by at least one tag which also occurs in the author descriptions of the pages.

We can see that, only a small portion of terms were applied by both users and authors for annotating and describing the same web pages. However, tags overlap more with keywords than description, even though description generally contains more terms. This may be because tags in nature are more likely to be keywords. They are the key terms chosen by users and authors to summarize the topics of web pages. Especially, a greater overlap exists between more popular tags and keywords. Table 2 list the 30 most
frequent tags and keyword terms used for the URLs in our data set.

In Table 4.1, the terms are ordered based on their frequency of usage. The bold terms are those used both by users as tags and by authors as keywords. We can see that more than half (17) of the 30 terms are shared. Especially, the eight most frequent keyword terms are all used as social tags. This indicates that a consensus to some extent exists between users and authors on the terms for describing web resources. However, different from keywords which only include topical terms, tags also contain some subjective and personal terms such as “fun”.

Table 4.1  30 most frequent tags and keyword terms

<table>
<thead>
<tr>
<th>Tags</th>
<th>Keywords</th>
</tr>
</thead>
<tbody>
<tr>
<td>reference, software, tools, design, web, shopping, free, programming, blog, art, music, news, resources, education, webdesign, technology, online, business, development, research, science, web2.0, internet, tutorial, games, computer, inspiration, how-to, fun, search</td>
<td>free, software, web, online, design, news, music, art, management, business, video, games, home, internet, books, education, digital, search, development, photography, computer, information, world, training, travel, school, windows, reviews, research, science</td>
</tr>
</tbody>
</table>

4.4 Clustering Analysis

An important function of metadata is to facilitate information organization (i.e., web clustering). As shown in the above section, a greater portion of tags and keyword terms are not present in the titles and content text of the pages. Although this indicates that both tags and author-provided keywords contain additional information beyond the page content, it is still unknown whether the additional information is useful for web
organization and searching. This section aims at answering this question by examining whether social tags and author-provided metadata are useful for enhancing the performance of web clustering. The popular K-means clustering method is adopted for such analysis.

4.4.1 Clustering Standard and Quality Metrics

Because there is no ground truth for web clusters, the 14 ODP top-level categories mentioned earlier is adopted as the gold standard for evaluating the clustering results. Cluster quality is evaluated by three metrics, F-score (Larson and Aone 1999), purity (Zhao and Karypis 2001), and normalized mutual information (NMI) (Strehl and Ghosh 2002). The F-score combines the information of precision and recall.

\[ F_{\text{score}} = \frac{2 \times \text{precision} \times \text{recall}}{\text{precision} + \text{recall}} \]

To compute purity, each cluster \( C_k \) is assigned to the category \( L_j \) which is most frequent in the cluster, and then the accuracy of the overall cluster assignments is measured by dividing the total number of correctly assigned documents by \( N \) (the number of clustered items).

\[ \text{Purity} = \frac{1}{N} \sum_k \max_j |C_k \cap L_j| \]

NMI is an increasingly popular measure of clustering quality. It is defined as the mutual information (I) between the cluster assignments C and a pre-existing class labeling L normalized by the arithmetic mean of the maximum possible entropies (H) of the empirical marginals, i.e.,

\[ \text{NMI}(X,Y) = \frac{I(X;Y)}{(\log k + \log c)/2} \]
A merit of NMI is that it does not necessarily increase when the number of clusters increases. All the three metrics range from 0 to 1; a higher metric value indicates better clustering quality.

4.4.2 Clustering Method and Results

In this study, K-means clustering method is adopted to compare the effectiveness of tag-based clustering, key-word based clustering, description-based clustering and content-based clustering. K-means is a simple but efficient and highly scalable clustering method. It iteratively calculates the cluster centroids and reassigns each document to the closest cluster until no document can be reassigned. Traditional K-means models documents with word vectors. In this study, web documents are not only modeled as the vectors of their content words but also as vectors of tags as well as vectors of keywords and description words. Specifically, the K-means clustering was run based on the following vector space models:

- **Word vector**: a web document is only represented by the vector of its content words. K-means, based on this vector model, is used as the baseline clustering method.
- **Tag vector**: a web document is only represented with the social tags applied to it.
- **Keyword vector**: a web document is only represented with the keyword terms used by the author to describe it.
- **Description vector**: a web document is only represented with the description words generated by the author.
• **(Word+Tag) vector**: Tags applied to a web document are viewed as additional words of the document, and combined with the original words of the document to form one vector.

• **(Word+Keyword) vector**: Keywords used to describe a web document are viewed as additional words of the document and combined with the original words of the document to form one vector.

• **(Word+Description) vector**: The words contained in the descriptions of a web document are viewed as additional words of the document and combined with the original words of the document to form one vector.

• **Word vector + Tag vector**: Each document is represented with two independent vectors: word vector and tag vector. During the clustering process, the distance from a document to a cluster centroid is calculated as the linear combination of the distance value based on word vector and the distance value based on tag vector.

• **Word vector + Keyword vector**: Each document is represented with two independent vectors: word vector and keyword vector. During the clustering process, the distance from a document to a cluster centroid is calculated as the linear combination of the distance values based on both word vector and keyword vector.

• **Word vector + Description vector**: Each document is represented with two independent vectors: word vector and description vector. During the clustering process, the distance from a document to a cluster centroid is calculated as the
linear combination of the distance values based on both word vector and description vector.

Another issue of K-means clustering is how to weigh the features. In this experiment, the tf-idf weighting scheme was adopted for all types of vector models. The tf-idf value of a term is decided by both the term frequency and its document frequency in the entire collection of documents. The cluster number of resources (web pages) was set to 14, equal to the number of ODP categories. Because the clustering algorithm relies on random initialization, each clustering process was run 10 times, and the mean of each quality metric across the 10 runs was used as the final score. For each run, the number of iterations was set to 20. The clustering based on word vector was viewed as the baseline. All words were lemmatized. Stop words and rare words with a document frequency less than 5 were filtered out.

Table 4.2 lists the clustering results of K-means based on different vector space models. Note that, in Table 4.2, ** indicates, compared to the baseline, the improvement is significant according to the paired-sample T-test at the level of p<0.01; * indicates the improvement is significant according to the paired-sample T-test at the level of p<0.05. The same symbols are applied in the following table.

From Table 4.2, we can see that, when clustering on a single vector, tag vector generates the best results. Compared to the baseline (clustering on word vector), clustering on either tag vector or keyword vector alone can improve the clustering performance significantly across all the quality metrics. However, when clustering on description vector alone, only the F-score is significantly improved and the NMI even decreases. This indicates that tags and keywords are more effective than description or
content words for distinguishing the topics of the web pages.

Table 4.2  The clustering quality of K-means on different vector space models

<table>
<thead>
<tr>
<th>Vector Space Model</th>
<th>F-Score</th>
<th>Purity</th>
<th>NMI</th>
</tr>
</thead>
<tbody>
<tr>
<td>word vector (baseline)</td>
<td>0.162</td>
<td>0.390</td>
<td>0.165</td>
</tr>
<tr>
<td>Tag vector</td>
<td>0.308**</td>
<td>0.444**</td>
<td>0.203**</td>
</tr>
<tr>
<td>keyword vector</td>
<td>0.303**</td>
<td>0.428**</td>
<td>0.191**</td>
</tr>
<tr>
<td>description vector</td>
<td>0.253**</td>
<td>0.398</td>
<td>0.160</td>
</tr>
<tr>
<td>(word+tag) vector</td>
<td>0.194</td>
<td>0.398</td>
<td>0.175</td>
</tr>
<tr>
<td>(word+keyword) vector</td>
<td>0.197</td>
<td>0.414**</td>
<td>0.194**</td>
</tr>
<tr>
<td>(word+description) vector</td>
<td>0.225**</td>
<td>0.406</td>
<td>0.182*</td>
</tr>
<tr>
<td>word vector + tag vector</td>
<td>0.307**</td>
<td>0.469**</td>
<td>0.249**</td>
</tr>
<tr>
<td>word vector + keyword vector</td>
<td>0.275**</td>
<td>0.461**</td>
<td>0.232**</td>
</tr>
<tr>
<td>word vector + description vector</td>
<td>0.226**</td>
<td>0.441**</td>
<td>0.210**</td>
</tr>
</tbody>
</table>

The inclusion of tags, keywords or descriptions as additional content also improves the clustering performance. However, the improvement is not significant. Especially, treating tags as additional words is least effective in improving the clustering performance.

Nevertheless, clustering on word vector+tag vector generates the best performance among all clustering models. It not only significantly improves the clustering quality over the baseline but also outperforms clustering on word vector+keyword vector and clustering on word vector +description vector.

Overall, the clustering results suggest that tags are most effective in enhancing web clustering when used as an independent information source. As for author-provided metadata, keywords in nature are similar to tags. However, keywords are less effective
than tags as an independent information channel for web clustering. This may be because keywords and page content are both created by the page authors. Descriptions generate better clustering results than keywords and tags only when they are used as the additional content of the web page. This suggests that descriptions are qualitatively most similar to page content compared to tags and keywords. In sum, the results show that social tags are the most reliable information source for identifying relevant web pages compared to either page content or author-provided keywords or descriptions.

4.5 Conclusion

In this chapter, the author investigates two approaches for metadata creation in the web environment: user-created tags and author-provided metadata. The user-created tags are social tags applied by users to annotate web pages. The author-provided metadata are the keywords and descriptions provided by the authors in the <meta> element placed in the head part of the pages’ HTML document.

In order to gain an insight into whether user-created tags and author-generated keywords contain additional information beyond the page content, the author examined their overlap with page title and text/body. The results prove that both tags and keywords add new information to existing page content. More than 50% of the tags and keywords are not present in the title and content of the pages. Compared to users, authors are more likely to use terms from the page content to annotate the pages. The author also compared tags with keywords and descriptions. Data analysis shows that users and authors only agree on a small portion of terms that can be used for describing the web pages.

To evaluate whether social tags or author-provided keywords and descriptions are
effective index terms or metadata which can used to enhance clustering and discovery of web resources, the author analyzed the data based on a clustering method (K-means) that incorporate the tags and keywords and descriptions into the clustering process. The results of K-means indicate that both tags and author-provided metadata can be leveraged to improve the clustering performance significantly, while tags are more effective than author-provided metadata as an independent information source for enhancing clustering quality. For future work, the author would like to investigate the effectiveness of user-created tags and author-provided metadata for other types of resources, such as videos and images.

4.6 Research Question Tested

*Question 1:* Are social tags effective document features which can be used to represent and index web documents in web mining and search applications?

In this chapter, the author investigated the additional information value provided by user-created social tags and author-provided metadata as well as their effectiveness in facilitating web clustering and discovery.
CHAPTER 5: CLUSTERING THE TAGGED WEB

5.1 Introduction

Compared to the metadata property, a more important property of social tagging is the tripartite network formed through users’ social tagging behaviors. The tripartite network of the social tagging system provides valuable information for ascertaining the topics of web documents, the semantics of tags and the domain interests of users.

Conventional clustering algorithms, such as k-means, cluster documents based on their textual content. Nevertheless, the content features of web documents are sometimes missing, misleading and unrecognizable due to the lack of well-controlled authoring styles and other reasons. Therefore, it is desirable to exploit other information sources to enhance clustering effectiveness. Links among documents have been used as important features for text mining in previous research. For instance, hyperlinks created by authors are viewed as evidence of an association between two web pages and utilized for ranking web search results (Kleinberg 1998, Page et al. 1998), classifying hypertext (Angelova and Weikum 2006, Li et al. 2008) and clustering web pages (Angelova and Siersdorfer 2006, Qi and Davison 2006). Besides the hyperlink network, the network formed in social tagging systems provides another rich information source for web mining. However, the tripartite structure of the social tagging network differs fundamentally from the well-studied hyperlink graph. In this paper, we examine how social tagging network can be exploited for web clustering.

The social tagging network has been well studied in previous research (Halpin et al.
2007, Lambiotte and Ausloos 2005, Maulik et al. 2010, Schenkel et al. 2008, Schmitz, et al. 2007, Wu et al. 2006). However, it has rarely been exploited for clustering purposes. This paper aims to explore whether and how the social tagging network can be utilized for improving web clustering. For this purpose, the author develops a novel clustering method based on the tripartite nature of the social tagging network. Different from conventional clustering approaches, which only cluster single type of data objects—the documents, the proposed clustering method clusters the three types of data objects (documents, users and tags) simultaneously. Accordingly, we call this method “Tripartite Clustering”. This clustering method fully relies on the link information within the social tagging network rather than the content of documents, thus it can be applied to any social tagging system. For instance, a large number of images and videos have been uploaded and annotated in social tagging systems such as flickr and Youtube. The Tripartite Clustering method can be directly used to cluster these image and video documents.

The proposed tripartite clustering approach is compared with two other clustering approaches based on the social tagging data. In the first approach, the tags applied to a document and the users who have annotated the document are viewed as additional features of the document and used to enrich the content-based document representation. K-means is used as the clustering method for this approach. In the second approach, tags and users are not directly used as document features for clustering, but viewed as bridges connecting related documents together. The cluster labels of the documents, which are connected to a document $d$ through shared tags and users, are used to help decide the cluster assignment of $d$. This approach is experimented with Link K-means method (Angelova and Siersdorfer 2006, Zhang et al. 2008). All three clustering methods were
experimented on the real-world social tagging dataset sampled from Delicious website. The clustering results were evaluated against a human-maintained web directory (Open Directory Project—ODP).

The organization of this chapter can be summarized as follows. Section 5.2 proposes the Tripartite Clustering model which clusters social tagging dataset based on the inherent network structure of social tagging systems. Section 5.3 introduces the social tag based K-means and Link K-means clustering methods. Section 5.4 presents the experimental settings and parameter tuning for the Tripartite Clustering method. Section 5.5 shows the clustering results of the three tag-based clustering approaches. Section 5.6 discusses and compares the performance of Tripartite Clustering with the other two tag-based clustering approaches. Finally, Section 5.7 concludes the chapter with major findings and future work.

5.2 Tripartite Clustering Model

In this section, we first introduce the novel Tripartite Clustering model on social tagging network.

5.2.1 Model Formulation

A social tagging network can be represented as a tripartite graph denoted by:

\[ G = (U, D, T, E^{(UD)}, E^{(UT)}, E^{(DT)}) \]  (5.1)

\( U, R, \) and \( T \) are finite sets of users, documents, and tags; \( E^{(UD)}, E^{(UT)}, \) and \( E^{(DT)} \) denote the three types of undirected links in the network: user-document, user-tag, and document-tag. \( E^{(UD)}, E^{(UT)}, \) and \( E^{(DT)} \) can be represented with three matrices, with row and column corresponding to two different types of nodes respectively.
In this graph, each type of node can be characterized by its links to the other two types of nodes. Specifically, a user’s interest can be featured by the documents he has previously annotated and the tags he has used; a document’s topic can be represented by the users who have annotated the document and the tags that have been assigned to the document. A tag’s semantics can be inferred from the documents that have been annotated with the tag and the users who have used this tag. Accordingly, a user $u_h$ can be represented with two vectors:

- $U^{(D)}_h = \{x^{(D)}_{hi} | i = 1,2,\ldots | D \}$, the document link vector, where $x^{(D)}_{hi}$ is the weight assigned to the link from user $u_h$ to document $d_i$, e.g. the frequency that document $d_i$ has been annotated by $u_h$, which normally equals to 1 in a typical social tagging system;

- $U^{(T)}_h = \{x^{(T)}_{hj} | j = 1,2,\ldots | T \}$, the tag link vector, where $x^{(T)}_{hj}$ is the weight assigned to the link from $u_h$ to tag $t_j$, e.g., the times that tag $t_j$ has been used by $u_h$.

Likewise, a document $d_i$ can be represented with:

- $D^{(U)}_i = \{y^{(U)}_{ih} | h = 1,2,\ldots | U \}$, $d_i$’s user link vector, where $y^{(U)}_{ih}$ is the weight assigned to the link from $d_i$ to $u_h$, and $y^{(U)}_{ih} = x^{(D)}_{hi}$.

- $D^{(T)}_i = \{y^{(T)}_{ij} | j = 1,2,\ldots | T \}$, $d_i$’s tag link vector, where $y^{(T)}_{ij}$ is the weight of the link from $d_i$ to $t_j$, e.g., the times that $d_i$ has been annotated with tag $t_j$.

And, a tag $t_j$ can be represented with:

- $T^{(U)}_j = \{z^{(U)}_{jh} | h = 1,2,\ldots | U \}$, $t_j$’s user link vector, where $z^{(U)}_{jh}$ is the weight of the link from $t_j$ to $u_h$, and $z^{(U)}_{jh} = x^{(T)}_{hj}$.

- $T^{(D)}_j = \{z^{(D)}_{ji} | i = 1,2,\ldots | D \}$, $t_j$’s document link vector, where $z^{(D)}_{ji}$ is the weight assigned to the link from $t_j$ to $d_i$, and $z^{(D)}_{ji} = y^{(T)}_{ij}$.
Based on this representation model, we can measure the similarity between any two nodes of the same type using their link distribution vectors. For instance, if two documents have similar user link distributions and similar tag link distributions, it can be inferred that these two documents have similar topics, and likewise for users and tags. Accordingly, the distance between two document nodes \( d_1 \) and \( d_2 \) can be the linear combination of the distance between their user link vectors \( dist(D^{(U)}_1, D^{(U)}_2) \) and the distance between their tag link vectors \( dist(D^{(T)}_1, D^{(T)}_2) \):

\[
dist(d_1, d_2) = \omega \cdot dist(D^{(U)}_1, D^{(U)}_2) + (1 - \omega) \cdot d(D^{(T)}_1, D^{(T)}_2)
\] (5.2)

In equation 5.2, \( \omega \) quantifies the influence of the user link vector and the tag link vector on document distance. The value of the distance can be calculated based on various similarity measures such as those proposed in (Markines et al. 2009). Distance metrics for user nodes and tag nodes can be formulated likewise.

Based on the representation model and distance metrics described above, the three types of nodes of the social tagging network can be clustered separately through any VSM (vector space model) based clustering method such as K-means, agglomerative hierarchical clustering or spectral clustering.

However, by clustering the three types of nodes separately, we ignore the iterative interactions among the different types of nodes. After analyzing a large amount of social tagging data, it is found that topically-related web pages are usually annotated with semantically-related tags by users with similar interests; and semantically-related tags are usually assigned to documents with similar topics and by users with similar interests; also, users with common interests usually annotate topically-related web pages using semantically-related tags. The same observation has also been made in (Bao et al. 2007,
Wu et al. 2006). Based on this observation, it is desirable to have a joint clustering approach which is capable of utilizing the interactions among the cluster structures of different types of nodes. This approach should be able to, for instance, cluster documents not only based on their direct links to the user nodes and tag nodes but also based on the cluster structures of the user nodes and the tag nodes. Therefore, even though two document nodes are not connected to the same users or same tags, they can still be assigned to the same cluster as long as they are connected to similar users (users with similar link features and belonging to the same cluster) or similar tags (tags with similar link features and belonging to the same cluster), as shown in Figure 5.1.

Figure 5.1 The similarity between two documents r1 and r2 are not only decided by their direct links to users, but also affected by the cluster structure of users, which is further decided by the links from users to tags.

In Figure 5.1, if the document nodes are clustered separately based on their direct links to users, then the two documents $d_1$ and $d_2$ would be assigned to different clusters inasmuch as they link to different user nodes and thus have zero similarity. However, if we take into account the link distributions from users to tags then we can find that user $u_1$ and $u_2$ are similar user nodes because they link to the same set of tag nodes. And $u_3$, $u_4$
and $u_5$ are also similar users. Accordingly, the similarity between $d_1$ and $d_2$ is no longer zero and they may be assigned to the same cluster.

5.2.2 Algorithm

In order to make use of the interactions among the cluster structures of different types of nodes, the author proposes an extended K-means algorithm based on the tripartite network of social tagging. The proposed algorithm takes the cluster numbers of documents, users and tags as input. Let’s denote the cluster number of documents, users, and tags with $k_R$, $k_U$, and $k_T$ respectively. Following the K-means approach, the algorithm first randomizes the cluster assignment of each node, and then iteratively calculates the centroid of each cluster and reassigns each node to the closest cluster based on its distance to the centroid of each cluster. In order to incorporate the interaction among the cluster structures of different types of nodes into the clustering process, the algorithm uses a novel metric to calculate the centroid of each cluster. In traditional K-means, the centroid of a cluster is the average of all the nodes in the cluster — that is, its coordinates are the arithmetic mean for each dimension separately over all the nodes in the cluster (MacKay 2003). In this proposed approach, the centroid of a cluster is decided not only by the features of the nodes in the cluster, but also by the cluster structures of other two types of nodes. For example, when clustering documents, the center of a document cluster is calculated based on the link features of the document nodes included in this cluster as well as the cluster structures of user nodes and tag nodes.

Let $C_m^{(D)}$ ($1 \leq m \leq k_d$) represents a document cluster, which includes $|C_m^{(D)}|$ documents. Each document can be represented with a user link vector and a tag link vector as described above. Let’s first consider the tag link vectors of these documents.
The value of the centroid vector of a document cluster $C_m^{(R)}$ at dimension $t_\tau$ is calculated as:

$$\text{centroid}^{(T)}_{m_\tau} = \frac{\sum_{d_i \in C_m^{(D)}, t_j \in C_n^{(T)}} y_{ij}^{(T)}}{|C_m^{(D)}| \ast |C_n^{(T)}|} \quad (t_\tau \in C_n^{(T)}) \tag{5.3}$$

where $C_n^{(T)}$ ($1 \leq n \leq k_T$) is the tag cluster that $t_\tau$ is assigned to. $t_j$ is any tag node included in cluster $C_n^{(T)}$, and $d_i$ is any document node assigned to $C_m^{(D)}$. $y_{ij}^{(T)}$ is the weight of the link from $d_i$ to $t_j$. According to equation (5.3), the value of a document cluster’s centroid vector at dimension $t_\tau$ not only depends on the links from $t_\tau$ to all the documents in the cluster, but also relies on the links from other tag nodes (those assigned to the same cluster as $t_\tau$) to the documents in the cluster. Note that the value of equation (5.3) can be viewed as an indicator of the association between the document cluster $C_m^{(D)}$ and the tag cluster $C_n^{(T)}$ (Long et al. 2006).

The superiority of the proposed method over traditional methods for calculating cluster centroid can be illustrated through the example shown in Figure 5.2. In Figure 5.2, $C_1^{(D)}$ is a document cluster composed by two documents $d_1$ and $d_2$. $t_1$: web2.0, $t_2$: folksonomy and $t_3$: socialbookmark are three tags assigned to tag cluster $C_1^{(T)}$. $t_4$: socialnetwork, $t_5$: facebook and $t_6$: socialsoftware are another three tags assigned to tag cluster $C_2^{(T)}$. Now, we want to cluster document $d_3$ based on its distance to all the document clusters including $C_1^{(D)}$. In $C_1^{(D)}$, because $d_1$ is annotated by $t_1$, $t_2$ and $t_4$, its tag link vector is $<1, 1, 0, 1, 0, 0>$, and $d_2$ is annotated by $t_1$, $t_4$ and $t_5$, so its tag link vector is $<1, 0, 0, 1, 1, 0>$. Likewise, the tag link vector of $d_3$ is $<0, 0, 1, 0, 0, 1>$. If calculated according to the traditional K-means approach, the centroid vector of $C_1^{(D)}$ is $<1, 0.5, 0, 1, 0.5, 0>$, and the cosine similarity between $d_3$ and the centroid of $C_1^{(D)}$ is 0. This is not a
proper result, because even though the tags used to annotate \( d_3 \) and \( d_1, d_2 \) in \( C_1^{(D)} \) are different, they are semantically related and assigned to the same tag clusters.

Now, if we consider the cluster structure of the tag nodes and calculate the centroid of the document cluster using (3), then we get the centroid vector of \( C_1^{(D)} \) as \(<1/3, 1/3, 1/3, 1/2, 1/2, 1/2>\). As a result, the cosine similarity between \( d_3 \) and the centroid of \( C_1^{(D)} \) changes to 0.566. Compared to 0, this is a more reasonable similarity value, because \( d_1, d_2 \) and \( d_3 \) are annotated by semantically related tags which have been assigned to the same tag clusters in previous iterations.

The centroid of a document cluster based on the user link vectors can be calculated likewise. The value of the centroid vector of a document cluster \( C_m^{(D)} \) at the user dimension \( u_\mu \) is calculated as:

\[
\text{centroid}_{m, \mu}^{(U)} = \frac{\sum_{d_i \in C_m^{(D)}, \lambda_i \in C_i^{(U)}} y_{i\mu}^{(U)}}{|C_m^{(D)}| \times |C_i^{(U)}|}, \quad (u_\mu \in C_i^{(U)})
\]  
(5.4)
, where $C_i^{(U)}$ ($1 \leq i \leq k_U$) is the cluster that $u_j$ belongs to. $u_j$ is any user node included in cluster $C_i^{(U)}$, and $d_i$ is any document node in $C_m^{(D)}$. $y_{ih}^{(U)}$ is the weight of the link from $d_i$ to $u_h$.

Like the distance between two documents in (2), the distance from a document $d_i$ to the centroid of a document cluster $C_m^{(D)}$ can be calculated as a linear combination:

$$dist(d_i, \text{centroid}_m) = \beta \cdot d(D_i^{(U)}, \text{centroid}_m^{(U)}) + (1 - \beta) \cdot d(D_i^{(T)}, \text{centroid}_m^{(T)})$$

(5.5)

, where $dist(D_i^{(U)}, \text{centroid}_m^{(U)})$ denotes the distance between $d_i$ and the centroid of $C_m^{(R)}$ based on the documents’ user link vectors, and $dist(D_i^{(T)}, \text{centroid}_m^{(T)})$ represents the distance based on the documents’ tag link vectors. $\beta$ quantifies the influence of a document’s user link vector on its cluster assignment. Greater value of $\beta$ means the cluster assignment of documents relies more on user link vectors and less on tag link vectors. The distance can be measured using various similarity metrics. In this paper, we adopt the cosine similarity.

The distance from a user (tag) node to the centroid of a user (tag) cluster can be calculated likewise as shown in (5.6) and (5.7).

$$dist(u_j, \text{centroid}_j) = \alpha \cdot dist(U_j^{(D)}, \text{centroid}_j^{(D)}) + (1 - \alpha) \cdot dist(U_j^{(T)}, \text{centroid}_j^{(T)})$$

(5.6)

$$dist(t_h, \text{centroid}_n) = \lambda \cdot dist(T_h^{(U)}, \text{centroid}_n^{(U)}) + (1 - \lambda) \cdot dist(T_h^{(D)}, \text{centroid}_h^{(D)})$$

(5.7)

In the above two equations, $\alpha$ decides to what extent the clustering of a user relies on its document link vector, and $\lambda$ specifies to what extent the cluster assignment of a tag depends on its user link vector.

Based on the proposed distance metrics, we develop an iterative clustering method called Tripartite Clustering, to discover the cluster structures of the social tagging network. Table 5.1 shows the algorithm of the proposed clustering method.
Table 5.1 Algorithm of Tripartite Clustering

Input:
The social tagging network $T_N = (U, D, T, E^{UD}, E^{UT}, E^{DT})$; Cluster numbers of document nodes, user nodes, and tag nodes: $k_D$, $k_U$, and $k_T$.

Output:
Cluster assignment of documents, users and tags: $C^{(D)}$, $C^{(U)}$, and $C^{(T)}$.

Method:
Initialize $C^{(D)}$, $C^{(U)}$, and $C^{(T)}$: assign each node to a random cluster;

Repeat:
   For each type of the node do
      Calculate the centroid of each document cluster as defined in equation (5.3) and (5.4). The centroids of user clusters and tag clusters are calculated likewise;
      For each document (user or tag) node do
         Calculate the distance between the node and the centroid of each document (user or tag) cluster according to equation (5.5), (5.6) or (5.7);
         Reassign the current node to the closest cluster based on the distance value.
   End For
End for

Until (The assignments no longer change OR Iteration Number $\geq$ Threshold)

5.3 Tag-based K-means and Link K-means

In tripartite clustering approach, web pages are clustered based on their user link features and tag link features with no dependence on document content. In this section, the author exploits social tagging for document clustering with two more clustering methods, K-means and Link K-means.
5.3.1 *K-means*

Traditional K-means represents documents with word vectors. In Chapter 4, in order to compare the metadata effectiveness of social tags and author-provided keywords and descriptions, the author experimented K-means clustering on a variety of tag-based vector models, keyword-based vector models and description-based vector models. In this section, focusing on the social tagging network, the author models the web documents with tag-based vectors and user-based vectors. Intuitively, if two documents are annotated with similar sets of tags, these two documents are likely to have similar topics. Similarly, documents annotated by the same users are more likely to be topically related than the documents annotated by different users. Specifically, we experiment with K-means on seven vector space models. The first four vector space models were also adopted in the clustering analysis of Chapter 4.

- **Word vector**: a web document is only represented with the vector of its content words.
- **Tag vector**: a web document is only represented with the social tags applied to it
- **(Word+Tag) vector**: Tags applied to a web document are viewed as additional words of the document, and combined with the original words of the document to form one vector.
- **Word vector + Tag vector**: Each document is represented with two independent vectors: word vector and tag vector.
- **User vector**: Each document is only represented with the users who annotated it.
- **Word vector + User vector**: Each document is represented with two independent vectors: word vector and user vector. During the clustering process, the distance from a
document to a cluster centroid is calculated as the linear combination of the distances calculated based on word vector and user vector.

- **Word vector + Tag vector + User vector**: Each document is represented with three independent vectors: word vector, tag vector and user vector. During clustering process, the distance from a document to a cluster centroid is calculated as the linear combination of the distances calculated based on these three vectors.

5.3.2 **Link K-means on Social Annotation**

On-page features are not always sufficient for text mining purposes. An approach to further enhancing clustering performance is to combine on-page features with features of the linked pages. Angelova and Siersdorfer (2006) developed such a method which incorporates the features of linked documents into the clustering process through iterative relaxation of clustering assignments. This algorithm can be built on top of any content-based clustering method. In our experiment, we build the algorithm on K-means, thus we call it Link K-means. The algorithm is briefly reviewed below.

Let \( D = \{d_i, \ i=1,2,\ldots,n\} \) be a document set, which is represented with an undirected graph \( G \). Each document \( d \in D \) corresponds to a node in the graph \( G \), and each link between two documents in \( D \) corresponds to an edge in \( G \). All the documents linked to document \( d \) is denoted by \( L(d) \). Let \( c(d) \) stand for the cluster of document \( d \), \( t(d) \) denotes the set of terms contained in \( d \). Then the cluster assignment of \( d \) is based on both \( d \)'s on-page features \( t(d) \) and the cluster distributions of its linked documents \( L(d) \). If \( \Phi_{i,d} \) denotes the probability of assigning a document \( d \) to cluster \( i \), then:

\[
\Phi_{i,d} = \Pr(c(d) = i \mid t(d), L(d)) = \Pr[c(d) = i \mid t(d), c(d_1), c(d_2), \ldots, c(d_k)]
\]
where $d_1$ through $d_k$ are the linked pages of document $d$ in $D$. If assuming that: (1) there is no direct coupling between the texts of a document and the cluster labels of its linked documents, and (2) the labels of a document’s neighbors are independent of each other, then we get the following formula:

$$
\Phi_{i,d} = \Pr[c(d) = i \mid t(d)] \sum_{c \in L(d)} \left( \prod_{d^* \in L(d)} \Pr[c(d) = i \land c(d^*) = j] \right)
$$

The above equation considers all combinations of the cluster assignments of $d$’s linked documents. For simplicity, for each of $d$’s linked document $d^* \in L(d)$, we only consider its maximum probable cluster assignment $c_{\max}(d^*)$. In addition, the link between $d$ and $d^*$ is weighted by $W$ based on some property of the link. Then the above equation can be simplified as:

$$
\Phi_{i,d} = \Pr[c(d) = i \mid t(d)] \prod_{d^* \in L(d)} \Pr[c(d) = i \land c(d^*) = c_{\max}(d^*)] * W
$$

The equation is resolved through iterative Relaxation Labeling techniques. First, the class label of each document is initialized through a content-based clustering process (K-means). Then the cluster assignment of each document is re-estimated using the label assignments of its neighbors and its own content. The re-estimation process based on the above equation iterates until the probability $\Phi_{i,d}$ for each document stabilizes or the changes drop below some threshold or the times of iterations reach a certain number.

The document links utilized in the algorithm can be various types, such as hyperlinks among web pages and co-authorship among academic papers (Zhang et al. 2008). Here, the author investigates whether the links formed in social tagging system can also be used in the same way to improve web clustering. The author examines three types of links existing among web pages in social tagging systems:
• **Co-user link:** if two web pages are annotated by the same user, then a co-user link is built among these two web pages. The number of users who have annotated both of these two web pages in our dataset is used as the weight of the link. By using this type of link in the link-based K-means algorithm, we assume that the cluster assignment of a web page can be influenced by the cluster distributions of the pages which are annotated by same users as the page.

• **Co-tag link:** if two web pages are annotated with the same tag, then a co-tag link is built between these two pages, the number of tags which are applied to both of these two pages is used as the weight of the link. By applying this type of link in the Link K-means algorithm, we assume that the cluster assignment of a web page is based on both its textual content and the cluster distributions of the pages which share the same tags as the page.

• **Co-user-tag link:** if two web pages are annotated by the same user with the same tag, then a co-user-tag link is created between these two pages. The frequency that these two pages are annotated with the same tag by the same user is the weight of the link. The author experiments on this type of link because the same user may not always annotate topically related web pages, and also the same tag may be used to annotate topically unrelated pages by different users. For instance, the tag “matrix” may be used by a user to annotate web pages about the mathematical references on matrices, while another user may use it to tag a web page about the movie “The Matrix”. Therefore, by applying this link in the Link-based K-means, we assume that only pages annotated with the same tag by the same user are topically related, and the cluster assignment of one of them can be influenced by the cluster assignments of the others.
Although both Tripartite Clustering and Link K-means rely on the linkage among documents, tags and users, they utilize the linkage in different ways. In Link K-means, tags and users are viewed as bridges connecting related documents together. Tags and users are not directly used as document features for clustering. Only the cluster labels of the documents, which are connected to a document $d$ through co-tag links, co-user links, or co-user-tag links, are used to help decide the cluster assignment of $d$. On the other hand, in the proposed Tripartite clustering, tags and users are viewed as inherent document features, and tags and users are clustered in synchrony with documents. The cluster assignments of documents are influenced by the cluster structures of tags and users. Besides, Tripartite Clustering fully relies on the network structure of social tagging, but Link K-Means is also based on the content features of documents.

5.4 Experiments

5.4.1 Dataset and Clustering Standard

The same dataset crawled from Delicious is used for experimentation. However, in order to reduce link sparsity, the URLs that were annotated only once and the tags that were used only once in the dataset were filtered out. For the rest of the URLs, the textual page content was crawled to support content-based clustering approaches.

Although all three types of nodes in the social tagging network are clustered through Tripartite Clustering, in our experiment we only evaluate the clustering results of web pages owing to the lack of ground truth for user clusters and tag clusters. Following the approach adopted in Chapter 4, the 14 top-level web categories of the Open Directory Project (ODP) is used as the standard of web page clusters. In the experiments section, all
the clustering results are evaluated against these 14 ODP categories. After overlapping the URLs from Delicious dataset with those classified under the 14 ODP web categories, the final dataset for experimentation contains 18509 URLs, 19452 tags and 4543 Users.

5.4.2 **Quality Metrics**

Cluster quality is evaluated by three quality metrics: F-score, purity, and normalized mutual information (NMI). All of them have been used in Chapter 4 for clustering analysis. All three metrics range from 0 to 1; a higher metric value indicates better clustering quality.

5.4.3 **Parameter Selection**

In the experiment with Tripartite Clustering, the distance between node and cluster centroid was measured with cosine similarity. Because Tripartite Clustering relies on random initialization, the author ran the algorithm 10 times and used the mean of each quality metric across the 10 runs as the final score. For each run, the number of iterations was set to 30. Like K-means, in Tripartite Clustering the number of clusters is an input parameter. Because Tripartite Clustering clusters documents, users, and tags all at once, how to choose appropriate $k_D, k_U$, and $k_T$ (cluster numbers of documents, users, and tags) is a problem. In the experiment, the clustering results of web pages were quantitatively evaluated against the 14 ODP categories, so $k_D$ was also set to 14. This setting is just for evaluation purposes and does not indicate that 14 is the best performing value of document cluster number. $k_U$ and $k_T$ were decided based on the clustering quality of a random sample of web pages. For different settings of $k_U$ and $k_T$, we measured the clustering quality of web pages on the small dataset. The settings that lead to best
clustering quality were selected. The experiments show that Tripartite Clustering achieved comparatively higher performance when \( k_U \) and \( k_T \) was set to be 25 and 35 respectively. Therefore, in following experiments, the author set \( k_U = 25 \) and \( k_T = 35 \).

Besides cluster numbers, Tripartite Clustering also involves three other parameters: \( \alpha \), \( \beta \) and \( \gamma \) in (5.5), (5.6) and (5.7) respectively. In order to select proper values for \( \alpha \), \( \beta \) and \( \gamma \), we varied each of them from 0.1 to 0.9 while setting the other two to be 0.5. The value which results in the best clustering result was selected. Figure 5.3, 5.4 and 5.5 display the plots of the clustering quality metrics of web pages on different values of \( \alpha \), \( \beta \) and \( \gamma \) respectively. In Figure 5.3, we can see that \( \alpha = 0.3 \) leads to the best clustering performance. Similarly, Figure 5.4 shows that the proposed clustering algorithm achieves the best performance when \( \beta = 0.5 \) and Figure 5.5 demonstrates that the optimal performing value of \( \gamma \) is also 0.5. Accordingly, we set \( \alpha = 0.3 \), \( \beta = 0.5 \), and \( \gamma = 0.5 \) respectively in following experiments.

Figure 5.3  The value of quality metrics for different values of \( \alpha \) when \( \beta = 0.5 \) and \( \gamma = 0.5 \). Greater value of \( \alpha \) means the cluster assignments of users rely more on their document link vectors and less on their tag link vectors.
5.5 Clustering Results

5.5.1 Tripartite Clustering

Although Tripartite Clustering generates cluster structures of web pages, tags, and users at the same time, only the quality of URLs was evaluated quantitatively against
ODP categories. In the future, we will explore more advanced approaches for evaluating the quality of tags clusters and user clusters.

<table>
<thead>
<tr>
<th>Cluster Method</th>
<th>F-Score</th>
<th>Purity</th>
<th>NMI</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tripartite Clustering</td>
<td>0.274</td>
<td>0.502</td>
<td>0.195</td>
</tr>
</tbody>
</table>

Table 5.2 shows the clustering quality of URLs. The quality of Tripartite Clustering is compared with that of other two clustering methods in following subsections. Although the clustering results of tags cannot be evaluated quantitatively because of the lack of ground truth, we can still get a general estimation about the quality of tag clusters by examining the semantics of tags assigned to each cluster. Table 5.3 displays the top 10 frequent tags from randomly selected 15 tag clusters. We can see that the top 10 tags in each cluster generally have similar or related semantics. The third column in Table 5.3 shows the dominant topic inferred from the semantics of the tags assigned to each cluster. For instance, tags in cluster 2 are generally related to web development, tags in cluster 6 are generally associated with web 2.0 or social network, and tags in cluster 13 generally deal with photography. The tag clusters can be applied to various purposes such as automatic tag recommendation and query expansion.
Table 5.3  Top 10 Frequent Tags of 15 Tag Clusters

<table>
<thead>
<tr>
<th>#tag</th>
<th>topic</th>
<th>top tags</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>teaching</td>
<td>teach elearn interactive child lesson parent algebra lessonplan literacy educational</td>
</tr>
<tr>
<td>2</td>
<td>web design</td>
<td>design webdesign cs inspiration flash html font usability icon typography</td>
</tr>
<tr>
<td>3</td>
<td>programming &amp; database</td>
<td>java library python database documentation apache perl sql source performance</td>
</tr>
<tr>
<td>4</td>
<td>software &amp; os</td>
<td>software opensource linux security network mac cm osx apple editor</td>
</tr>
<tr>
<td>5</td>
<td>outdoor activity</td>
<td>car camp buy cycle gear auto outdoors bargain transportation survival</td>
</tr>
<tr>
<td>6</td>
<td>web 2.0 &amp; social network</td>
<td>web2.0 bookmark twitter share socialsoftware mashup feed 2.0 competition facebook</td>
</tr>
<tr>
<td>7</td>
<td>science &amp; education</td>
<td>science education write learn math literature podcast mathematics psychology astronomy</td>
</tr>
<tr>
<td>8</td>
<td>reference</td>
<td>reference search research travel map history searchengine archive health dictionary</td>
</tr>
<tr>
<td>9</td>
<td>web development</td>
<td>program development javascript ajax php webdev framework ruby code script xml</td>
</tr>
<tr>
<td>10</td>
<td>news &amp; entertainment</td>
<td>news fun humor politics funny movy daily entertainment comicstrip film webcomic</td>
</tr>
<tr>
<td>11</td>
<td>news</td>
<td>newspaper usa international mlf documentary press temp canada favorite translate</td>
</tr>
<tr>
<td>12</td>
<td>shopping</td>
<td>shop architecture gadget home fashion clothe store museum toy tshirt</td>
</tr>
<tr>
<td>13</td>
<td>photography</td>
<td>photo photography graphic image photoshop wallpaper stock 3d sound desktop</td>
</tr>
<tr>
<td>14</td>
<td>music &amp; multimedia</td>
<td>music audio tv mp3 radio stream television guitar record player</td>
</tr>
<tr>
<td>15</td>
<td>art</td>
<td>art cool magazine animation idea creativity retro style draw</td>
</tr>
</tbody>
</table>

5.5.2  K-means on Social Tagging Network

An issue of K-means clustering is how to weight the document features. For word features and tag features, two weighting functions, tf (term frequency) and tf·idf were adopted. The tf·idf value of a term is decided by both its term frequency and its document frequency in the whole collection. As for the user features, because a document can be
annotated by a user at most once, the value of a document’s user vector at each dimension can only be 1 or 0.

The experiment used the same configuration as the Tripartite Clustering: the cluster number K was set to 14; the iteration times was set to 30; each evaluation was run 10 times with random initialization and the average was used as the final clustering result. All the content words were lemmatized. Stop words and the words with document frequency less than 5 were filtered out.

Table 5.4 lists the results of K-means clustering based on different vector space models. ** indicates, compared to the baseline (K-means on Word vector), the improvement was significant according to the paired-sample t-test at the level of p<0.01; * indicates the improvement was significant according to the paired-sample t-test at the level of p<0.05. These two symbols are applied in all the following tables.

From Table 5.4, we can see that, K-means based on *tag vector* and *word vector+tag vector* generates the best results among the vector space models without using user vectors. They both significantly outperform the baseline method (K-means on word vector) under either tf or tf·idf scheme. *(word+tag) vector* also improves the clustering performance compared to the baseline, but the improvement is not so significant, especially under the tf weighting scheme. This confirms that tags provide a different information channel from words for web clustering. In fact, from the clustering results we can conclude that tags are more reliable features than words for web clustering. Besides, for all the vector space models involving tags, the clustering based on tf·idf weighting always outperforms the clustering based on tf weighting. It indicates that tf·idf is more appropriate for weighting tags.
Table 5.4 Clustering Quality of K-Means with Different Vector Space Models and Two Weighting Schemes

<table>
<thead>
<tr>
<th>Vector Space Model</th>
<th>F-Score</th>
<th>Purity</th>
<th>NMI</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>tf</td>
<td>tf·idf</td>
<td>tf</td>
</tr>
<tr>
<td>Word vector</td>
<td>0.123</td>
<td>0.103</td>
<td>0.424</td>
</tr>
<tr>
<td>(Word+Tag) vector</td>
<td>0.144</td>
<td>0.168**</td>
<td>0.426</td>
</tr>
<tr>
<td>Tag vector</td>
<td>0.252**</td>
<td>0.292**</td>
<td>0.505**</td>
</tr>
<tr>
<td>Word vector + Tag vector</td>
<td>0.266**</td>
<td>0.278**</td>
<td>0.512**</td>
</tr>
<tr>
<td>User vector</td>
<td>0.154</td>
<td>-</td>
<td>0.404</td>
</tr>
<tr>
<td>Word vector + User vector</td>
<td>0.175*</td>
<td>0.162**</td>
<td>0.433</td>
</tr>
<tr>
<td>Word vector + Tag vector + User vector</td>
<td>0.263**</td>
<td>0.298**</td>
<td>0.522**</td>
</tr>
</tbody>
</table>

User features are not as effective as tags for enhancing document clustering. The K-means solely relying on user vector even perform worse than the baseline method according to Purity and NMI. Although combining user vector with word vector improves the clustering performance to some extent, the improvement is not substantial. However, if the tags are also introduced into the clustering process as in the word vector + tag vector + user vector model, the clustering performance is remarkably improved. As shown in Table 5.4, in half of the cases, the word vector + tag vector + user vector model leads to the best clustering performance among all vector space models.

5.5.3 Link K-means on Social Tagging Network

The clustering performance of Link K-means can be influenced by the number of links or bridges built among the documents in the collection. In order to control the numbers of the three types of links at the same scale, the author set different threshold to the weight of the three types of links. For the co-user link, the threshold was set to 2,
which result in 2,285,080 co-user links. For the co-tag link, the threshold of the weight was set to 10, which generates 2,720,850 co-tag links. For co-user-tag link, the threshold of the weight was set to 1, which produces 2,358,696 co-user-tag links. Like Tripartite clustering and K-means, for each type of link, the author ran the algorithm 10 times and took the average as the final clustering result. The iteration number was set to 30 and the cluster number was still set to 14. The content word was weighted using both tf and tf·idf scheme. Table 5.5 shows the clustering results.

### Table 5.5  Clustering Quality of Link-Based K-Means Based on Different Link Types

<table>
<thead>
<tr>
<th>Link Type</th>
<th>F-Score</th>
<th>Purity</th>
<th>NMI</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>tf</td>
<td>tf·idf</td>
<td>tf</td>
</tr>
<tr>
<td>Co-user link</td>
<td>0.157</td>
<td>0.230</td>
<td>0.452</td>
</tr>
<tr>
<td>Co-tag link</td>
<td>0.149</td>
<td>0.213</td>
<td>0.439</td>
</tr>
<tr>
<td>Co-user-tag link</td>
<td><strong>0.181</strong></td>
<td><strong>0.243</strong></td>
<td><strong>0.472</strong></td>
</tr>
</tbody>
</table>

Note that Link K-means using the co-user-tag link evinces the best performance across all quality metrics under both tf and tf·idf weighting. It significantly outperforms both Link K-means using the co-user link and Link K-means using the co-tag link. This result confirms that the same user may annotate web pages with different topics; the same tag can likewise be applied to topically different web pages by different users. The same user, however, would always apply the same tag to web pages with similar topics.
5.6 Discussion

The purpose of this chapter is to explore approaches of improving web clustering by leveraging the social tagging network. The experiments with three different clustering methods (Tripartite Clustering, K-means, and Link K-means) all prove the value of social tagging network for web clustering but from different perspectives. Table 5.6 displays the results of Tripartite Clustering and the best cluster models of K-means and Link K-means. K-means on word vector is used as the baseline method.

It can be seen that all three clustering methods improve the performance of the baseline method significantly by incorporating the social tagging data into the clustering process through different approaches. K-means on \textit{word vector + tag vector + user vector} treats tags and users as additional document features and linearly combines them with word vector to determine the distance between a document and a cluster centroid. In Link K-means, users and tags act as bridges which bring topically related web pages together. The cluster of a document is determined not only by its content words but also by the cluster assignments of neighboring documents. Tripartite Clustering fully relies on the relationships among the web pages, users and tags for clustering. It clusters the three entities simultaneously; the cluster structure of one entity can influence the cluster structure of the other two entities.

Both Tripartite Clustering and K-means on the \textit{word vector + tag vector + user vector} significantly outperforms Link K-means with co-user-tag link across all quality metrics. K-means on \textit{word vector + tag vector + user vector} also outperforms Tripartite Clustering in most cases, especially when the tf-idf weighting scheme is used. However, the difference is not significant based on the t-test. Therefore, it can be concluded that
when applied to clustering documents, the Tripartite Clustering method, which fully relies on the social tagging network, can compete with clustering methods which rely on both content and social tagging.

Table 5.6  Comparison of the Clustering Quality of Different Clustering Methods

<table>
<thead>
<tr>
<th>Clustering Method</th>
<th>F-Score</th>
<th>Purity</th>
<th>NMI</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Tf</td>
<td>tf·idf</td>
<td>tf</td>
</tr>
<tr>
<td>K-means on Word vector</td>
<td>0.123</td>
<td>0.103</td>
<td>0.424</td>
</tr>
<tr>
<td>K-means on Word vector +Tag vector +User vector</td>
<td>0.263**</td>
<td>0.298**</td>
<td>0.522**</td>
</tr>
<tr>
<td>Link K-means with Co-user-tag link</td>
<td>0.181**</td>
<td>0.243**</td>
<td>0.472**</td>
</tr>
<tr>
<td>Tripartite Clustering</td>
<td>0.274**</td>
<td>-</td>
<td>0.502**</td>
</tr>
</tbody>
</table>

Apart from document clusters, Tripartite Clustering also outputs the cluster structures of users and tags as well as the relationships among the clusters of different entities. The information can be exploited for various applications, such as automatic tag recommendation and personalized search. Moreover, because Tripartite Clustering does not rely on the content of annotated documents, it can be utilized to cluster non-textual documents annotated in some social tagging systems. For instance, we would like to use it to cluster the images in the Flickr datasets in the future.

Tripartite Clustering can also compete with K-means and Link K-means in terms of processing time. If the iteration number is \( I \), the number of clusters is \( K \), the number of entities to be clustered is \( N \), the dimensionality of an entity’s feature space is \( D \), and the number of documents linked to a document is \( L \), then the time complexity of K-means, Link K-means and Tripartite Clustering is \( O(IN) \), \( O(IND + INL) \), and
O(\(IKND+IK^2N\)) respectively. We can see that the computation complexities of all three clustering algorithms are at the same scale. In real experiments, Tripartite Clustering takes longer, because it clusters three types of entities all at the same time. If we use K-means to cluster each type of entity individually by transforming the links into features for each type of node, the total computational time would be close to the time used for Tripartite Clustering.

5.7 Conclusion and Future Work

This chapter explores how to utilize social tagging network for document clustering. The author extends two content-based cluster methods, K-means and Link K-means, to incorporate the social tagging information during the clustering process. The author also proposes a novel clustering method called Tripartite Clustering to discover the cluster structures of the tripartite network of social tagging. The clustering methods were applied to a real-world social tagging dataset and evaluated against a human-edited web directory. Experimental results show that all the social tagging-based clustering methods significantly outperform the content-based K-means in clustering web pages. The author also compared the Tripartite Clustering method with K-means which uses content words, social tags and users together as document features, and Link K-means which clusters web pages based on both on-page content and the links formed through social tagging network. Experimental results show that the content-independent Tripartite Clustering method significantly outperforms Link K-means but performs slightly worse than the K-means approach.

In the future, the proposed Tripartite Clustering method can be further improved in several ways. Firstly, the feature weighting scheme used in the method needs to be
refined. Experimental results of K-means and Link K-means both show that the clustering quality of these two methods can be further improved by weighting document features (content words and tags) with the \( \text{tf-idf} \) scheme. In the future, the author would like to explore more advanced weighting schemes for Tripartite Clustering.

A limitation of Tripartite Clustering is that each object can only be assigned to one cluster. In other words, like K-means, Tripartite Clustering is a hard clustering algorithm. Its clustering results cannot correctly reflect the documents with multiple topics, or social tags with multiple semantics, or users with more than one interest. In the future, the author would like to develop a soft clustering method based on the Tripartite Clustering algorithm by representing the cluster membership of each object with probability values.

Although the proposed method also clusters users and tags, the author could not evaluate the quality of user and tag cluster structures owing to the lack of ground truth. In the future, the author would like to develop both quantitative and qualitative approaches for evaluating the clustering results of users and tags. Finally, the author also plans to experiment with the proposed clustering method on datasets collected from other social tagging systems, especially those for image and video annotation.

5.8 Research Questions Tested

**Question 2:** How to enhance web clustering based on the social tagging network?

This chapter explores how to utilize social tagging network for document clustering. A novel clustering method called Tripartite Clustering is developed to discover the document cluster structures from the tripartite social tagging network.
CHAPTER 6: PERSONALIZED SEARCH ON TAGGED WEB

6.1 Introduction

The social tagging system not only provides a channel to learn the topics of web resources, but also provides a window to identify users’ information needs and preferences. In this chapter, the author investigates how to realize and enhance personalized search based on the social tagging network.

In a social tagging system, we can learn user interests from users’ social tagging history, for instance based on the tags he/she ever created or the documents he/she ever annotated. Although the application of social tagging for personalization seems to be straightforward, it is still a challenge to fully exploit the social tagging network for personalized search. A major reason is that, different from previously well-studied network structures (such as the hyperlink network, author-document network, or document-word network), the social tagging network is composed by the tripartite relations among users, tags and web documents.

As pointed out by Lambiotte and Ausloos (2005), it is always possible to project the tripartite network of social tagging into bipartite or unipartite networks, and then apply existing methods for bipartite and uniparite relationships to model and exploit the social tagging network. For instance, a user-document-tag tripartite relation can be projected into three undirected bipartite relations: user-tag relation, user-document relation and document-tag relation. However, the projection inevitably causes information loss.

Tensors have been proved to be effective for modeling the tripartite relationship
involved in social tagging data (Rendle et al. 2009, Wetzker et al. 2009). However, tensor based algorithms (such as high order singular value decomposition) are always limited in efficiency and scalability. Considering the rapidly growing tagging community and tagged content, as well as the sparsity related to the power-law distribution in the social tagging systems, highly scalable and efficient methods are desirable for modeling and applying the tripartite network of social tagging data. Especially, when we apply social tagging data for web search, where instant response to users’ queries is a key factor affecting users’ experience and evaluation for the systems, efficiency is more essential.

In this chapter, the author proposes an efficient and effective personalized search framework (TripleQE) based on the tripartite network of social tagging. In other words, the proposed search framework utilizes query expansion for personalization and ranking. However, different than previous expansion methods which only expand the query keywords, the proposed search framework also expands the queries in the user dimension. More importantly, it incorporates the tripartite relationship of social tagging for query expansion. Based on two datasets collected from real-world social tagging systems, we compared our personalized search framework with two language model (LM) based search models and a personalized search method, called FolkRank, which employs a page-rank like algorithm on the social tagging graph to identify important documents for a user and a tag query. The experimental results demonstrate that our personalized search framework can outperform all the benchmark methods in terms of effectiveness. It is also much more efficient than FolkRank for personalized search.

The rest of the chapter is organized as follows: Section 6.2 introduces the proposed personalized search framework. Section 6.3 discusses the approaches of incorporating
language model based search models into the proposed framework for document ranking. Section 6.4 specifies the experiments used for evaluation. Section 6.5 presents and discusses the experimental results. Finally, section 6.7 summarizes the whole chapter.

![Tripartite social tagging hyper-graph](http://example.com/tgraph.png)

Figure 6.1 An example of the tripartite social tagging hyper-graph. Each time a user annotates a web page with a tag, a tripartite hyper-edge (triple) is created among the user, the resource and the tag. This figure contains eight hyper-edges: $e_1, e_2, e_3, e_4, e_5, e_6, e_7, e_8$.

### 6.2 Personalized Search Framework

In this section, we present the personalized search framework developed on the social tagging network. In existing literature, personalized search primarily adopts two strategies. One strategy is re-ranking, which re-ranks the documents returned by a non-personalized search system using previously constructed user profiles. The other strategy is query expansion which modifies the original query by expanding it with other terms to incorporate user information in the searching process. The proposed personalized search framework adopts a strategy similar to query expansion. However, in our approach, the
expansion is not only based on query terms, but also based on the query users and the tripartite relationship among tags, users, and documents.

6.2.1 Model Formulation

Before introducing the personalized search method, we first define a graph model to represent the social tagging network. A social tagging network can be represented by a 3-uniform hyper-graph denoted as:

\[ \text{HG} = (U, D, T, E^{(UDT)}) \]  \hspace{1cm} (6.1)

- \( U, R, \) and \( T \) are finite sets of user vertices, document vertices, and tag vertices;
- \( E^{(UDT)} \) represents a set of hyper-edges. Each hyper-edge is a triple of a user node, a document node and a tag node: 
  \[ E^{(UDT)} = \{e = \{u, d, t\} \mid f(u, d, t) = 1\} \]

\( f(u, d, t) \) is a function which equals to 1 if tag \( t \) is used by user \( u \) to annotate document \( d \) in the dataset; otherwise, it equals 0. Namely, \( f(u, d, t) \) defines whether \( u, d, t \) forms a triple in the social tagging hyper-graph.

For instance, the social tagging network represented in Figure 1 contains eight tripartite hyper-edges: 
\[ E^{(UDT)} = \{e_1, e_2, e_3, e_4, e_5, e_6, e_7, e_8\} = \{\{u_1, \text{filezilla-project.org, filezilla}\}, \{u_1, \text{filezilla-project.org, client}\}, \{u_2, \text{filezilla-project.org, filezilla}\}, \{u_2, \text{filezilla-project.org, ftp}\}, \{u_2, \text{filezilla-project.org, server}\}, \{u_2, \text{www.01ftp.com, ftp}\}, \{u_3, \text{www.01ftp.com, server}\}, \{u_3, \text{www.01ftp.com, web2ftp}\}\} \]

Note that the graph representation of the social tagging system proposed here is different from the one proposed in equation 5.1 of chapter 5, where the ternary user-document-tag hyper-edge is projected into three types of binary edge: user-tag edge, user-document edge and document-tag edge.
Given a query \( q=(u_q, t_q) \), where \( u_q \) is the user who proposes the query and \( t_q \) is the query term, the purpose of the personalized search method is to return a ranked set of documents based on the hyper graph defined above. Here, we only consider the queries which only contain a single query term, but the proposed framework can be easily extended to support complicated queries composed by more than one term. Besides, we also assume that the query term is one of the tags contained in the dataset. This assumption is based on two considerations. Firstly, in the social tagging system, users assign tags to documents, because they can later use the tags as keywords to search the documents. Thus the tags can be considered as query keywords to some extent. Secondly, the tagged documents represent the web resources that are most interesting to online users. Therefore, it is very likely that the tags assigned to the documents comprise of popular search keywords. The empirical study conducted by Heymann et al. (2008) shows that there is a significant overlap between popular query terms and tags.

Now if the query term is a tag, the query \( q = (u_q, t_q) \) is actually part of a triple in the social tagging graph \( HG \). Given the whole hyper-graph of social tagging, a simple personalized search system can just return the set of documents that have been annotated by user \( u_q \) with tag \( t_q \): \( \{ d \mid f(u_q, d, t_q) = 1 \} \). However, such documents may not be available. The user who proposes the query may have not tagged any documents with the query term yet. Besides, even if the user did not apply the tag \( t_q \) to a document, it does not necessarily mean that the user would not think \( t_q \) a proper tag for the document, considering that the number of proper tags that a user can come up with during the tagging process is always limited. A practical and effective personalized search system should be flexible enough to be able to return a satisfying number of relevant documents
in these conditions. An approach to address this issue is query expansion. Different from traditional query expansion method in which the query terms are expanded without taking into account the user factor, here we propose a new query expansion approach which also considers the user information in the query. The proposed approach is based on three assumptions:

**Assumption 1**: The user $u_q$ would think $t_q$ a proper tag for document $d$, if document $d$ has been annotated by $u_q$ with some tags similar to $t_q$.

This assumption is intuitively true. For instance, in figure 6.1, if we know server is a similar tag to ftp based on their co-occurrence in documents, since user $u_3$ has tagged document www.01ftp.com with server, it is reasonable to assume that $u_3$ would also think it proper to annotate document www.01ftp.com with ftp. Based on this assumption, we can expand the query tag $t_q$ with its similar tags that have been used by user $u_q$. Then the documents which appear in the same triple with $u_q$ and tags similar to $t_q$ can be returned as relevant documents to $q = (u_q, t_q)$. Formally, these documents can be denoted as $D_1$:

$$D_1 = \{ d \mid f(u_q, d, t_i) = 1, (t_i \in T_q) \} \quad (6.2)$$

where $T_q$ is the set of tags similar to $t_q$. The expansion with Assumption 1 is based on the tag dimension. In the personalized search framework, we also propose to expand the query from the user dimension based on Assumption 2.

**Assumption 2**: The user $u_q$ would think $t_q$ a proper tag for document $d$, if the users similar to $u_q$ have annotated document $d$ with $t_q$.

For example, in Figure 1, if we know that $u_1$ and $u_2$ are similar users based on their tagging history, it is reasonable to assume that $u_1$ would also think ftp a proper tag for
filezilla-project.org since \(u_2\) has applied \(ftp\) to filezilla-project.org. Based on Assumption 2, we can expand the query \(q = (u_q, t_q)\) with users that are similar to \(u_q\) and connected to \(t_q\). Then the documents which have been tagged with \(t_q\) by users similar to \(u_q\) can be returned as relevant documents to \(q = (u_q, t_q)\). Formally, the set of returned documents based on Assumption 2 can be represented as:

\[
D_2 = \{d \mid f(u_j, d, t_i) = 1, (u_j \in U_q)\} \quad (6.3)
\]

, where \(U_q\) is the set of users similar to \(u_q\).

Finally, based on Assumption 1 and Assumption 2, we can further derive Assumption 3.

**Assumption 3:** The user \(u_q\) would think \(t_q\) a proper tag to annotate document \(d\), if users similar to \(u_q\) have annotated \(d\) with tags similar to \(t_q\).

For instance, in Figure 1, if we know \(server\) is a similar tag to \(ftp\) based on their co-occurrence in documents, and \(u_1\) and \(u_2\) are similar users based on their tagging history, we can assume that \(u_1\) would think \(server\) a proper tag for filezilla-project.org based on the fact that \(u_2\) has applied \(ftp\) to filezilla-project.org. Through **Assumption 3**, we implement personalized search with two-dimensional expansions: user expansion and tag expansion. These two expansions are achieved simultaneously based on the social tagging hyper-graph. Specifically, the documents which have been annotated by the users similar to \(u_q\) with tags similar to \(t_q\) are returned as relevant documents to \(q = (u_q, t_q)\). Formally, these documents can be denoted as document set \(D_3:\)

\[
D_3 = \{d \mid f(u_j, d, t_i) = 1, (u_j \in U_q, t_i \in T_q)\} \quad (6.4)
\]

\(U_q\) and \(T_q\) have the same definition as specified in Equation (6.2) and (6.3).
6.2.2 Triple-based Query Expansion Method

After knowing how the query should be expanded, the author proposes a simple but very effective and efficient search method: Triple-based query expansion search method (TripleQE), which identifies the documents included in $D_1$, $D_2$ and $D_3$ within a unified framework.

Given a query $(u_q, t_q)$, for each of the triples contained in the social tagging hyper-graph, we first calculate its similarity score to the query according to equation (6.5):

$$
SimScore(u_j, t_i) = \alpha \cdot \text{sim}(t_i, t_q) + (1 - \alpha) \cdot \text{sim}(u_j, u_q)
$$

(6.5)

$u_j$ and $t_i$ are the user node and tag node of a triple satisfying $f(u_j, d, t_i) = 1$. $\alpha (0 \leq \alpha \leq 1)$ decides to what extent the user information influences the search results. When $\alpha = 1$, the query is expanded at the tag dimension. In this case, the search system returns documents that are annotated by the tags similar to the query tag without considering the user information. Whereas, when $\alpha = 0$, the query is expanded at the user dimension. In this case, the search system returns documents that are tagged by users similar to the query user without taking into account the tag information.

In equation (6.5) the similarity between two users $\text{sim}(u_j, u_q)$ can be calculated based on the percentage of the documents or tags shared by them in the dataset, or based on their group membership. The similarity between two tags $\text{sim}(t_i, t_q)$ can be calculated with many methods as introduced by Markines et al. (2009). In our current implementation, the similarities between two users and two tags are calculated based on the user-document relation and tag-document relation respectively.

Specifically, each user is represented with a vector $\vec{U}$ with $|D|$ dimensions. $|D|$ is the
number of document nodes in the social tagging graph.

\[ \bar{U}_j = \langle w_k \mid k = 1,2,\ldots,|D| \rangle \]

\(w_k\) is the normalized frequency that a user is connected with document \(d_k\) through a tripartite hyper-edge in the social tagging graph.

\[ w_k = \frac{c(u_j ; d_k)}{\sum_k c(u_j ; d_k)} \]

, where \(c(u_j ; d_k)\) is the number of times that user \(u_j\) and document \(d_k\) appear in the same triple in the social tagging hyper-graph.

Based on the vector representation, the similarity between two users is measured with the cosine similarity of their corresponding vectors.

Likewise, each tag is also represented with a vector.

\[ \bar{T}_i = \langle v_k \mid k = 1,2,\ldots,|D| \rangle \]

\(v_k\) is the normalized frequency that the tag is connected with document \(d_k\) through a tripartite hyper-edge in social tagging hyper-graph.

\[ v_k = \frac{c(t_i ; d_k)}{\sum_k c(t_i ; d_k)} \]

, where \(c(t_i ; d_k)\) is the number of times that tag \(t_i\) and document \(d_k\) co-occur in the same triple in the entire social tagging graph. Then the similarity of two tags is also measured with the cosine similarity of their corresponding vectors.

After calculating the SimScore for each triple in the graph, then the documents contained in the \(M\) triples with highest SimScore are used as candidate documents to return. We use \(TR^M\) to denote the set of \(M\) triples with highest SimScore,
\[ TR^M := \arg \max_{\{u, d, t\} \in E^{UDT}} \ \text{sim}[(u_j, t_i), (u_q, t_q)] \]  \hspace{1cm} (6.6)

Then the set of candidate documents can be denoted with \( D_c \):

\[ D_c = \{ d \mid \{u_j, d, t\} \in TR^M \} \hspace{1cm} (6.7) \]

Only considering documents contained in the top \( M \) triples has two advantages: firstly, it helps filter out the noisy information and thus improve search precision; secondly, it helps reduce the data dimension and thus improves search efficiency.

It is easy to see that \( D_c \) in fact consists of documents from \( D_1 \), \( D_2 \) and \( D_3 \) defined in Equation (6.2), (6.3) and (6.4). For instance, in equation (6.6), when \( t_i \) and \( t_q \) are the same tag (i.e., \( \text{sim}(t_i, t_q)=1 \)) and \( 0<\text{sim}(u_j, u_q)<1 \), the documents connected to \( u_j \) and \( t_i \) belongs to \( D_1 \); when \( u_j \) and \( u_q \) are the same user (i.e., \( \text{sim}(u_j, u_q)=1 \)) and \( 0<\text{sim}(t_i, t_q)<1 \), the documents associated with \( u_j \) and \( t_i \) belongs to \( D_2 \); when \( 0<\text{sim}(u_j, u_q)<1 \) and \( 0<\text{sim}(t_i, t_q)<1 \), the documents connected to \( u_j \) and \( t_i \) belongs to \( D_3 \).

In the final step, each candidate document in \( D_c \) is ranked based on two factors: the number of times that the document appears in \( TR^M \), and the \( \text{SimScore} \) of the triple in which it occurs. Formally, given the social tagging hyper-graph \( G \) and the query \((u_q, t_q)\), each candidate document \( d \) is ranked based on score \( S_G(d) \):

\[ S_G(d) = \sum_{\{u_j, d, t\} \in TR^M} \text{sim}[(u_j, t_i), (u_q, t_q)] \]  \hspace{1cm} (6.8)

Because the score \( S_G(d) \) is calculated fully based on the social tagging hyper-graph, we call it graph score.
6.3 Enhanced Ranking

The method TripleQE proposed above fully depends on the social tagging graph for personalized searching without considering document content. This makes it flexible enough to be directly applied for searching different types of tagged collections, such as web pages, articles, images, videos, etc. However, when the content of the tagged document is available, it is still favorable to utilize content information during the searching process, because document content always contains useful information for document search as demonstrated in traditional information retrieval (IR) models. Therefore, we propose a method to incorporate the content information into the personalized search framework during the ranking step.

Besides $S_G(d)$ calculated based on the social tagging hyper-graph, for each candidate document included in $D_c$, we also calculate a content-based score based on a traditional information retrieval model. In our experiments, we adopt the language model-based IR method to calculate the content-based score $S_{LM}(d)$.

$$S_{LM}(d) = P(t_q \mid d)$$

$P(t_q \mid d)$ can be calculated with different smoothing methods as proposed by Zhai and Lafferty (2001). Here we adopt the Bayesian smoothing method (Zhai and Lafferty 2001).

$$S_{LM}(d) = P(t_q \mid d) = \frac{tf(t_q; d) + \mu^* p(t_q \mid C)}{\sum_w tf(w; d)} \quad (6.9)$$

where $tf(t_q; d_k)$ is the term frequency of the query term $t_q$ in document $d$ as a content word. $p(t_q \mid C)$ is the maximum likelihood estimate of $t_q$ in the entire document collection.
\( \mu \) is the Dirichlet prior. We can see that the value of \( S_{LM}(d) \) is independent of the query user.

Finally, each candidate document in \( D_c \) is ranked based on the linear combination of the two scores \( S_G(d) \) and \( S_{LM}(d) \):

\[
S(d) = \lambda \cdot S_G(d) + (1 - \lambda) \cdot S_{LM}(d)
\]

(6.10)

where \( \lambda (0 \leq \lambda \leq 1) \) decides to what extent the rank of a document is influenced by \( S_G(d) \). When \( \lambda = 1 \), the rank of a candidate document in the returned list fully relies on the graph-based score \( S_G(d) \), whereas when \( \lambda = 0 \), the rank of a candidate document in the returned document list is fully decided by the language model-based score \( S_{LM}(d) \).

In our experiments, besides the original words contained in the document, the tags that have been applied to a document are viewed as another type of document content. This is because some tagged web resources, such as images or videos, do not contain textual content. In these cases, we can use tags as substitute content. Therefore, depending on whether document content is represented with words or tags, the \( S_{LM}(d) \) of a document can have different values. When tags are used as to represent document content, the content score is calculated as:

\[
S_{LM}(d) = P'(t_q \mid d) = \frac{tf'(t_q; d) + \mu \cdot p'(t_q \mid C_i)}{\sum_t tf'(t; d)}
\]

(6.11)

where \( tf'(t_q; d) \) is the term frequency of the query term \( t_q \) in document \( d \) as a tag. \( P'(t_q \mid C_i) \) is the maximum likelihood estimate of \( t_q \) in the entire document collection represented with tags.
6.4 Experiments

This section introduces the dataset on which the experiments were conducted, the methods and performance metrics for evaluation, the baseline methods for comparison and the approaches for parameter setting.

6.4.1 Datasets

In order to evaluate the performance of the proposed personalized search method, the author prepared two real-world dataset from two social tagging systems: Delicious and Bibsonomy. The Delicious dataset is a sample of the same dataset used in chapter 4 and 5. It contains 189,910 social bookmarks and composes 41,190 web pages by 4414 users with 28,740 unique tags. For each web page, the author also crawled its textual content from the web. The Bibsonomy dataset was downloaded from its data dump page (http://www.kde.cs.uni-kassel.de/bibsonomy/dumps). The data dump created by Jan 01, 2011 is used. In Bibsonomy, the tagged documents include both web pages and published articles. In order to study whether the proposed method is effective for different document types, the author only selected the posts created for publications in Bibsonomy dataset. Besides, in order to assess the effectiveness of the content-based ranking, and compare the proposed personalized search method with content-based IR, the author only chose those publications with abstract information available in its corresponding BibTeX entries. The title and the abstract of each publication are used as its word content. The final Bibsonomy dataset for experiment contains 79,733 publications, 143,863 tags, 34,649 users.

Both dataset, especially Bibsonomy, have data sparsity problem due to the data
nodes in the “long tail”. Hence, in order to increase data density and restrict the
evaluation to the “dense” part of the data, for each dataset, the author applied the
algorithm proposed by Hotho et al. (2006) to generate a $p$-core in which each document,
user and tag occurs in at least 2 posts. The $p$-core of Delicious data at level 2 contains
628,318 triples, 4,184 users, 16,910 documents, and 11,270 unique tags. The $p$-core of
Bibsonomy data at level 2 contains 93,943 triples, 648 users, 8,966 documents and
10,590 tags.

6.4.2 Evaluation Method

Evaluation of personalized search is always challenging due to the lack of benchmark
queries and relevant results. For personalized search, only the querying user can judge
whether a document is relevant to his/her query. Therefore, the notion of relevance for
personalized search is highly subjective and dependent on the querying user. Previous
studies of personalized search primarily do evaluation through user experience studies
(Chirita et al. 2005, Chirita et al. 2007, Qiu and Cho 2006, Teevan et al. 2005) and search
engine query logs (Sun et al. 2005). However, user studies are costly in terms of time and
efforts, and the search engine query log is always limited in availability. In this paper, we
adopt the evaluation method proposed by Xu et al. (2008), which uses user’s tagging
behavior to infer user’s relevance judgment. In other words, if the querying user $u$
annotated a document $d$ with a tag $t$, this user would think $d$ is a relevant document when
he/she issues a query contains $t$. The same evaluation method has been used in other
personalized search studies and proved to be effective (Harvey et al. 2011, Schenkel et al.
2008).

Based on this evaluation method, for each dataset we create a set of pseudo queries.
Each query is a randomly selected user-tag pair \((u_q, t_q)\) which has appeared in some triples in the dataset. All the documents which have co-occurred with a query user-tag pair in a triple are considered as the relevant documents to this query. In other words, given a query \(q = (u_q, t_q)\), we consider the union of all the documents which have been annotated by user \(u_q\) with tag \(t_q\) as the ground truth for query \(q\). The triples that contain the pseudo queries are selected out as the test data, and the remaining triples are used for training. In this way, we make sure that all the query tags created by the corresponding query users are removed from its original posts.

For Delicious dataset, 4000 pseudo queries were created, and for Bibsonomy data, 200 pseudo queries were created. For each dataset, we separate its pseudo queries and the associated ground truth documents into two subsets with equal size. One subset is used for parameter tuning; the other subset is used for final evaluation.

### 6.4.3 Evaluation Metrics

Two metrics are used to measure the search effectiveness of the proposed method and comparative methods. The first metric is MAP (mean average precision), which is a widely-used evaluation metric in IR community. It considers the order in which the returned documents are presented and favors ranking relevant documents higher.

\[
MAP = \frac{\sum_{q=1}^{O} AveP(q)}{Q}
\]

The second metric is Success@k, which is the ratio of times when at least one relevant document is returned in the top \(k\) documents. For both metrics, the values at rank 1, rank 5, rank 10 and rank 20 are reported.
6.4.4 Parameter Settings

In the proposed Triple-based Query Expansion (TripleQE) method, there are two important parameters to be set: $\alpha$ in equation (6.5) specifying the influence of user information on the search results, and $M$ in Equation (6.6) denoting the number of triples from which the candidate documents are generated. The values of these two parameters were decided one by one based on the testing subset for parameter tuning. First, we fixed the value of $\alpha$ and test $M$ on a large range of settings from 40 to 200. $M$ was set to the value which leads to the best performance. It is found that when $M$ was set to be 80 for Delicious data, and 60 for Bibsonomy data, the proposed TripleQE generated the best result. Note that here the documents were only ranked based on the social tagging graph score $S_G(d)$.

After setting $M$ to the optimized value, we varied the value of $\alpha$ from 0.0 to 1.0 to identify the best setting. Also the value leading to the best performance was used in the final evaluation.

![Figure 6.2 MAP@10 over different setting of $\alpha$](image)
Figure 6.2 and Figure 6.3 respectively show the MAP and Success value at rank 10 when $\alpha$ was set to different values from 0.0 to 1.0. We can see that for both dataset, when $\alpha$ equals to 0.5, MAP@10 and Success@10 reach the highest value. Besides, we can see that when queries were only expanded at the user dimension ($\alpha = 0$), the performance is worst. At the other extreme case, when $\alpha = 1$ or the queries were merely expanded at the query tag dimension, the MAP@10 and Success@10 are also relatively low. This indicates that both user information and tag information can be utilized to improve the search performance, and the search performance can be most effectively improved when the query is expanded at both user dimension and tag dimension.

Besides $\alpha$ and M, if using the enhanced ranking methods proposed in section 6.3, it is also necessary to decide the value of the Dirichlet Prior $\mu$ in Equation (6.9) and (6.11), as well as $\lambda$ in equation (6.10) which determines to what extent the rank of a document is influenced by $S_G (d)$ or $S_{LM} (d)$.

The Dirichlet prior was set in an independent search system only based on the
language model with Bayesian smoothing (BayesLM). The value that leads to the best performance for BayesLM is adopted in the ranking algorithm proposed in section 6.3. It is found that for tag based language models, when $\mu$ was set to be 300 for Bibsonomy data and 400 for Delicious data, the BayesLM reached optimal performance. However, in word based language models, for both datasets when $\mu$ was set to be 0, the Bayesian LM generates best results. This may be because the query keywords used for experiments are tags themselves, thus they are in nature different from the content words. Accordingly, the smoothing strategy which is used to assign non-zero probability to unseen words is ineffective for tag-based queries.

In order to identify the optimal value of $\lambda$, we also varied it from 0.0 to 1.0. The value which produces the best search results was used for final evaluation. Figure 6.4 shows MAP@10 over different values of $\lambda$ when word-based BayesLM is combined with TripleQE for ranking. We can see that, for Delicious data, when $\lambda=0.7$, the searching performance is the best; and for Bibsonomy data, when $\lambda=0.3$, the searching performance is optimized. Since $(1-\lambda)$ indicates the influence of word-based BayesLM for document ranking, it means that word information should be more emphasized in Bibsonomy data than in Delicious data during the searching process. This maybe because, compared to the word content of the Bibsonomy documents, which are publication abstracts, the word content of the Delicious documents, which are web pages, contains more noisy information.

Figure 6.5 displays MAP@10 over different values of $\lambda$ when tag-based BayesLM is combined with TripleQE to rank the candidate documents. It shows that when $\lambda$ is set to be 0.8 for Bibsonomy data and 0.9 for Delicious data, the searching performance is
optimized. The settings of different parameters used in final evaluation are summarized in Table 1.

Figure 6.4 The value of MAP@10 over different settings of $\lambda$ when word-based language model with Bayesian smoothing is incorporated into TripleQE to rank the candidate documents.

Figure 6.5 The value of MAP@10 over different settings of $\lambda$ when tag-based language model with Bayesian smoothing is incorporated into TripleQE to rank the candidate documents.
6.4.5 Baselines

The proposed personalized search method is compared with three baseline methods. Two methods are non-personalized search method based on the Language Model with Bayesian smoothing (Zhai and Lafferty 2001). One method is personalized search method based on FolkRank proposed in (Hotho et al. 2006).

**BayesLM_word**: word based language model with Bayesian Smoothing, where the document content is represented with words. This is the traditional Language Model with Bayesian Smoothing.

\[
P(t_q | d) = \frac{tf(t_q; d) + \mu \cdot p(t_q | C)}{\sum_{w} tf(w; d)}
\]

However, as mention in parameter settings section, because when Dirichlet prior \( \mu \) is set to be 0, this method generates the best result. In the final evaluation experiment, \( \mu \) was set to be 0, which equals to non-smoothed language model.

**BayesLM_tag**: Tag based language model with Bayesian Smoothing, in which document content is represented with tags.

\[
P(t_q | d) = \frac{tf'(t_q; d) + \mu \cdot p'(t_q | C_i)}{\sum_{t} tf'(t; d)}
\]

In our experiment, the Dirichlet prior \( \mu \) was optimized to 300 for Bibsonomy dataset, and 400 for Delicious dataset.

**FolkRank**: FolkRank is an algorithm inspired by PageRank. It is proposed by Hotho (2006) for document search and ranking based on the social tagging network. The key idea of FolkRank is that the tag which is used to annotate important documents by important users is important itself. The same holds for documents and users. Therefore the social tagging system can be represented with a hyper-graph in which the vertices
mutually reinforce each other by spreading their weights. Specifically, a weighted adjacency matrix \( A \) is constructed with each element corresponding to the co-occurrence of two nodes in the social tagging graph. The weight of each node is updated as \( w \leftarrow dA\!w\!+\!(1-d)p \), where \( p \) is a preference vector, and \( d \) is the damping factor. First, a global weight \( w^0 \) is generated for each node by setting all the entries of \( p \) to 1. Then an adapted weight vector \( w^1 \) is obtained by setting the entries corresponding to the query user and query tag to \(|U|\) and \(|D|\) respectively. Finally each nodes are ranked according to \( w = w^1 - w^0 \). The top ranked document nodes are returned as the relevant documents to the query. In our experiment, the damping factor \( d \) was set to be 0.7.

### Table 6.1 Searching methods used for experiments

<table>
<thead>
<tr>
<th>Denotation</th>
<th>Searching Methods</th>
<th>Parameter Settings</th>
</tr>
</thead>
<tbody>
<tr>
<td>BayesLM_word</td>
<td>Word based language model with Bayesian Smoothing</td>
<td>( \mu = 0 )</td>
</tr>
<tr>
<td>BayesLM_tag</td>
<td>Tag based language model with Bayesian Smoothing</td>
<td>( \mu = 400 ) (Delicious) ( \mu = 300 ) (Bibsonomy)</td>
</tr>
<tr>
<td>FolkRank</td>
<td>FolkRank</td>
<td>( d = 0.7 )</td>
</tr>
<tr>
<td>TripleQE</td>
<td>Triple-based query expansion</td>
<td>( \alpha = 0.5 ) ( M = 80 ) (Delicious) ( M = 60 ) (Bibsonomy)</td>
</tr>
<tr>
<td>TripleQE+LM_word</td>
<td>TripleQE combined with BayesLM_word based ranking</td>
<td>( \lambda = 0.7 ) (Delicious) ( \lambda = 0.3 ) (Bibsonomy)</td>
</tr>
<tr>
<td>TripleQE+LM_tag</td>
<td>TripleQE combined with BayesLM_tag based ranking</td>
<td>( \lambda = 0.9 ) (Delicious) ( \lambda = 0.8 ) (Bibsonomy)</td>
</tr>
</tbody>
</table>

### 6.5 Results

This section discusses the experimental results of the proposed personalized search method and the baseline methods introduced above. All the metric values reported here
were generated using the testing subset for final evaluation. Table 6.1 lists the denotations of all the search methods and the settings of their parameters in the final experiments.

Table 6.2   MAP of different searching methods on Delicious dataset

<table>
<thead>
<tr>
<th>Method</th>
<th>MAP</th>
<th>@5</th>
<th>@10</th>
<th>@20</th>
</tr>
</thead>
<tbody>
<tr>
<td>BayesLM_word</td>
<td>0.1225</td>
<td>0.1736</td>
<td>0.1803</td>
<td>0.1846</td>
</tr>
<tr>
<td>BayesLM_tag</td>
<td>0.2599</td>
<td>0.3582</td>
<td>0.3678</td>
<td>0.3735</td>
</tr>
<tr>
<td>FolkRank</td>
<td>0.2699</td>
<td>0.3819</td>
<td>0.3976</td>
<td>0.4081</td>
</tr>
<tr>
<td>TripleQE</td>
<td>0.2745</td>
<td>0.4181</td>
<td>0.4408</td>
<td>0.4559</td>
</tr>
<tr>
<td>TripleQE+LM_word</td>
<td>0.2957</td>
<td>0.4243</td>
<td>0.4448</td>
<td>0.4665</td>
</tr>
<tr>
<td>TripleQE+LM_tag</td>
<td>0.2993</td>
<td>0.4316</td>
<td>0.4549</td>
<td>0.4700</td>
</tr>
</tbody>
</table>

Table 6.3    Success of different searching methods on Delicious dataset

<table>
<thead>
<tr>
<th>Method</th>
<th>Success</th>
<th>@1</th>
<th>@5</th>
<th>@10</th>
<th>@20</th>
</tr>
</thead>
<tbody>
<tr>
<td>BayesLM_word</td>
<td>0.145</td>
<td>0.298</td>
<td>0.342</td>
<td>0.422</td>
<td></td>
</tr>
<tr>
<td>BayesLM_tag</td>
<td>0.320</td>
<td>0.566</td>
<td>0.618</td>
<td>0.698</td>
<td></td>
</tr>
<tr>
<td>FolkRank</td>
<td>0.346</td>
<td>0.632</td>
<td>0.720</td>
<td>0.826</td>
<td></td>
</tr>
<tr>
<td>TripleQE</td>
<td>0.404</td>
<td>0.718</td>
<td>0.828</td>
<td>0.902</td>
<td></td>
</tr>
<tr>
<td>TripleQE+LM_word</td>
<td>0.368</td>
<td>0.724</td>
<td>0.824</td>
<td>0.898</td>
<td></td>
</tr>
<tr>
<td>TripleQE+LM_tag</td>
<td>0.420</td>
<td>0.755</td>
<td>0.840</td>
<td>0.935</td>
<td></td>
</tr>
</tbody>
</table>

Table 6.2 and Table 6.3 list the MAP and Success rate of different methods on Delicious data. From both tables, we can see that the two non-personalized searching methods BayesLM_word and BayesLM_tag have lowest searching performance
compared to other personalized search methods. This indicates that the searching performance can be improved by incorporating user factors into the searching process. The tag based BayesLM performs much better than word based BayesLM. This can be caused by two reasons. Firstly, the query keywords are selected from tags in our experiment. Therefore, the document language model estimated based on the tags is more accurate for the query-likelihood ranking method. Secondly, document tags contain less noisy information than the word content. This indicates that, even in non-personalized search models, the tagging information can be exploited to improve the searching performance.

Among personalized search methods, all the TripleQE based search methods perform better than FolkRank. Moreover, they are also much more efficient than FolkRank. This is because, the FolkRank algorithm needs to be run online, and for every query it has to traverse each binary edge in the social tagging graph for several iterations, which makes its response time extremely longer than other algorithms. For the TripleQE algorithm, the similarities between two tags or two users can be calculated offline. Thus, for each query, it only needs to go through the ternary hyper-edges in the social tagging graph once, and the number of the ternary hyper-edges is much smaller than that of the binary edges in the social tagging graph.

Although the proposed TripleQE already performs better than all the baseline methods, it shows that the searching performance can be further improved by introducing the language model based methods into the ranking stage. TripleQE+LM_tag has the best performance among all the searching methods on Delicious data. TripleQE+LM_word performs better than TripleQE but worse than TripleQE+LM_tag. This indicates that, for
Delicious data, tag-based BayesLM is more effective than word-based LM when combined with Triple QE for ranking.

Table 6.4 MAP of different searching methods on Bibsonomy dataset

<table>
<thead>
<tr>
<th>Method</th>
<th>@1</th>
<th>@5</th>
<th>@10</th>
<th>@20</th>
</tr>
</thead>
<tbody>
<tr>
<td>BayesLM_word</td>
<td>0.0683</td>
<td>0.1245</td>
<td>0.1400</td>
<td>0.1467</td>
</tr>
<tr>
<td>BayesLM_tag</td>
<td>0.1591</td>
<td>0.2742</td>
<td>0.2926</td>
<td>0.3018</td>
</tr>
<tr>
<td>FolkRank</td>
<td>0.1932</td>
<td>0.3151</td>
<td>0.3411</td>
<td>0.3565</td>
</tr>
<tr>
<td>TripleQE</td>
<td>0.2159</td>
<td>0.3321</td>
<td>0.3703</td>
<td>0.3983</td>
</tr>
<tr>
<td>TripleQE+LM_word</td>
<td>0.2562</td>
<td>0.4131</td>
<td>0.4454</td>
<td>0.4722</td>
</tr>
<tr>
<td>TripleQE+LM_tag</td>
<td>0.1778</td>
<td>0.3539</td>
<td>0.3819</td>
<td>0.4117</td>
</tr>
</tbody>
</table>

Table 6.5 Success of different searching methods on Bibsonomy dataset

<table>
<thead>
<tr>
<th>Method</th>
<th>@1</th>
<th>@5</th>
<th>@10</th>
<th>@20</th>
</tr>
</thead>
<tbody>
<tr>
<td>BayesLM_word</td>
<td>0.133</td>
<td>0.311</td>
<td>0.455</td>
<td>0.556</td>
</tr>
<tr>
<td>LM_tag</td>
<td>0.320</td>
<td>0.520</td>
<td>0.623</td>
<td>0.693</td>
</tr>
<tr>
<td>FolkRank</td>
<td>0.380</td>
<td>0.600</td>
<td>0.740</td>
<td>0.800</td>
</tr>
<tr>
<td>TripleQE</td>
<td>0.380</td>
<td>0.780</td>
<td>0.870</td>
<td>0.960</td>
</tr>
<tr>
<td>TripleQE+LM_word</td>
<td>0.450</td>
<td>0.790</td>
<td>0.900</td>
<td>0.960</td>
</tr>
<tr>
<td>TripleQE+LM_tag</td>
<td>0.390</td>
<td>0.800</td>
<td>0.830</td>
<td>0.950</td>
</tr>
</tbody>
</table>

Table 6.4 and Table 6.5 show the MAP and Success of different searching methods on Bibsonomy data. The performance of different searching methods displays similar patterns on Bibsonomy data and Delicious data, except that for Bibsonomy data the TripleQE+LM_word outperforms TripleQE+LM_tag and has the best performance among all methods. In Bibsonomy data, the tagged documents are publications, and the
content words are collected from the titles and abstracts of published articles, which usually contain high quality information. But in Delicious dataset, the documents are web pages with content words crawled from the Web which has little authority control. Therefore, the word content of Bibsonomy documents is more informative and less noisy than the word content of Delicious documents. Based on this finding we can conclude that, when combined with the TripleQE for document ranking, word-based BayesLM is more effective than tag-based BayesLM when the word content contains high quality information. This is probably because, unlike tags whose information has already been used in TripleQE, words represent an additional and independent information source for the TripleQE method. This also confirms the value of document content in the searching process.

In summary, based on the discussion above, we can conclude that the proposed personalized search method (TripleQE) is effective for enhancing the performance of personalized search. It outperforms some competitive search models like the Language model with Bayesian smoothing. It also performs better than the well-defined personalized search model FolkRank in both effectiveness and efficiency. Moreover, when content-based LM is incorporated into the TripleQE for document ranking, the performance can be further improved.

6.6 Conclusion

The social tagging behavior by online users not only creates useful information to learn the topics of tagged web documents, but also provides a window to learn a user’s information needs and interests. In this chapter, the author explores how to improve
personalized search by exploiting the social tagging data. For this purpose, a personalized search method is developed based on the social tagging hyper-graph which is formed by the ternary hyper-edges among users, tags, and documents. The proposed method implements personalized search through query expansion strategy. During the searching process, a query is expanded by utilizing the user relation identified through users’ common interests in documents, the tag relation based on tags’ co-occurrence in documents, and most importantly, the tripartite relation among users, tags and document. Experiments on two real-world datasets collected from Delicious and Bibsonomy demonstrates that the proposed personalized search framework is more effective than baseline search methods. Moreover, when the proposed method is combined with language model based search methods, the performance can be further improved. In the future, the author would like to experiment the proposed personalized search framework on datasets of much larger scale and compare it with more search methods.

6.7 Research Question Tested

*Question 3: How to exploit the social tagging network for personalized web search?*

This chapter develops a personalized search method (TripleQE) based on the social tagging hyper-graph. The proposed method implements personalized search by expanding queries based on the tripartite relation among users, tags and document.
CHAPTER 7: A TOPIC-PERSPECTIVE MODEL FOR THE SOCIAL TAGGING SYSTEM

7.1 Introduction

In order to utilize the social tagging network for web mining, it is prerequisite to properly model the relationships among the different entities involved in the social tagging systems, including users, documents, tags, and content words of documents. In previous two chapters, the social tagging network is modeled as a graph, in which the relations among the three entities (document, user and tag) are represented with graph edges, and each entity is represented as vectors of the other two entities. Although the graph model is effective for tag-based clustering and search applications, it has some limitations. First, the vector representations of the entities based on the graph model can have very high dimension, which makes related calculations inefficient. Secondly, the model does not reflect the real social tagging process, because the edges among documents, users and tags are undirected and unordered. Moreover, the role of document content is not reflected in the graph model. In this chapter, the author adopts a different approach to model the social tagging system. Specifically, the author attempts to model the interactions among different entities involved in the social tagging system by simulating the generation process of social tags with a probabilistic generative model. The proposed model can be used to discover the topical structures of documents and the perspectives of users. Most importantly, it can help identify the influence of document topics and users’ personal perspectives on the generation of tags.

So far, a variety of probabilistic generative models have been developed to model
the social tagging system. For instance, Wu et al. (2006) propose a probabilistic generative model in which the three entities in a social tagging system (tags, resources, and users) are mapped to a common conceptual space, which is represented by a multi-dimensional vector with each dimension corresponding to a knowledge category. Besides, hierarchical Bayesian models using Probabilistic Latent Semantic Analysis (PLSA) and Latent Dirichlet Allocation (LDA) have also been proposed to model the social tagging process (Kashoob et al. 2009, Plangprasopchok and Lerman 2007, Ramage et al. 2009). In this chapter, the author proposes a new hierarchical Bayesian model based on the well-known LDA for simulating the generation process of social tagging.

In the proposed model, the author uses two different latent variables to represent resource topics and user perspectives. Here the perspective of a user not merely refers to the user’s interest, but also affected by the user’s expertise, motivation, language and other personal factors. First, resource topics are generated and represented with word distributions. Then the identified topics of a resource are used to generate the tags together with user perspectives. Accordingly, the model is named as “Topic-Perspective (TP) Model”.

A distinct feature of Topic-Perspective model is that, during the tag generation process, resource topics and user perspectives together generate the social tags for a resource, and each tag differs in the extent of depending on resource topics or user perspectives. The rationality of this design is evidenced by the existence of social tags with various functional purposes. Recall the seven different functions of tags proposed by Golder and Huberman (2006) (see section 1.2 and section 2.1.3). Sen et al. (2006) summarize the seven tag functions proposed by Golder and Huberman into three
categories: factual tags, subjective tags, and personal tags. Intuitively, Factual tags are more closely related to resource content and extrinsic to the taggers, while the Subjective tags and Personal tags are less connected to resource topics and more influenced by users’ perspectives. For instance, on Delicious, the URL (http://www.brainyquote.com/) titled “Famous Quote and Quotations at BrainyQuote” is annotated by factual tags like quotes, quotations, writing, literature, etc. Meanwhile it is also annotated with the subjective and personal tags such as funny, cool, interesting, etc. Factual tags are more valuable than subjective and personal tags for representing resource content, and thus are more effective when used as additional information in web mining and search tasks. Therefore, it is necessary to identify whether a tag relies more on resource topics or user factors. Based on the proposed generative model for social annotation, we are able to estimate the probabilistic that each tag is generated from user perspectives or resource topics.

Moreover, different from other models, in this model, the tag generation process is modeled separately from the content word generation process. This design is based on the observation that tags are generated differently from content words: the content words contained in a document are generated by a single or a small group of authors sharing common interests, while the social tags of a document are generated by many users with different perspectives. A set of tags applied to a resource can reflect both users’ perspectives and the resource’s topics.

Based on this model, we can learn the topical distribution of each document, the perspective distribution of each user, the word distribution of each topic, the tag distribution of each topic, and the tag distribution of each user perspective, and the
probabilistic of each tag being generated from resource topics or user perspectives.

The rest of this chapter is organized as follows: Section 7.2 presents the proposed TP Model for social annotation and introduces the parameter estimation process. Section 7.3 evaluates the performance of the proposed model and compares it with other two models described in previous research. Section 7.4 discusses possible applications of the proposed model. Section 7.5 concludes this chapter.

7.2 Topic-Perspective Model

This section introduces the proposed Topic-Perspective Model for social annotation. This model depicts the social annotation process and the generation process of content terms in a unified framework. The motivation behind this model is to represent and connect all the observed and hidden variables in a unified framework. By estimating this model, we can learn the topical structure of the documents, terms, and tags, the tagging perspectives of users and the representation of user perspectives with tags at the same time.

7.2.1 Model Formulation

The design of this model reflects the real social tagging process. Before a tag was created by a user for a document, the document terms already exist. Therefore, we first consider the term generation process which can be modeled with a standard LDA model. Then when a user annotates a document, two factors act on the tag generation process. One is the topics of the document; the other is the perspective adopted by the user. Even for the tags created by the same user, the extent that each tag is affected by user’s perspectives may be different, because, as mentioned, each tag can be created for different functional purposes. For instance, some tags are created to specify the topics of the resources, while
other tags may be created for self-reference, quality evaluation, and opinion expression. Intuitively, the tags of the former kind are more dependent on the topics of the documents, while the latter kind are more affected by users’ personal perspectives.

Figure 7.1 The Topic-Perspective model for social tagging

Figure 7.2 Graphical representation of (a) CI-LDA and (b) CorrLDA for modeling social annotations
In order to reflect this nature of social annotations in the generative model, we adopt a switch variable to control the influence of user perspectives and document topics on tag generation. The proposed model is illustrated in Figure 7.1. The meanings of notations used in Figure 7.1 are summarized in Table 7.1.

As shown in Figure 7.1, this model primarily comprises of two parts split by the dash line. The right part is essentially the standard LDA, which models the generation of content terms contained in the documents. For each word \( w \) in a document \( d \), a topic \( z \) is first sampled, and then the word \( w \) is drawn conditioned on the topic.

The document \( d \) is generated by repeating the process \( N_d \) times, which is the number of word tokens in \( d \). The left part of Figure 7.1 models the generation of tags. Each tag \( t \) created by user \( u \) for document \( d \) can be drawn from either the topics associated with \( d \)’s content words or \( u \)’s perspectives. To decide the source of each tag, a switch variable \( x \) is introduced. The value of \( x \) (which is 0 or 1) is sampled based on a binomial distribution \( \lambda \) (with a Beta prior \( \gamma \)). When the sampled value of \( x \) equals 1, tag \( t \) is drawn from the topic \( z_t \), which is uniformly sampled from the topics learned from the words in document \( d \). The red arrows in Figure 2 show this process. When \( x \) equals 0, a perspective \( p \) is first sampled from the perspective distribution \( (\theta_u) \) for user \( u \), and then the tag \( t \) is drawn from the tag distribution \( \psi_p \) of perspective \( p \). The blue arrows in Figure 7.1 illustrate this procedure. Overall, the generation process of words and tags in the Topic-Perspective model can be described as follows:

1) For each of the D documents \( d \), sample \( \theta^{(d)}_d \sim \text{Dirichlet} (\alpha_d) \);

2) For each of the U users \( u \), sample \( \theta^{(u)}_u \sim \text{Dirichlet}(\alpha_u) \);
3) For each of the K topics k, sample $\phi^{(w)}_k \sim \text{Dirichlet}(\beta_w)$, and sample $\phi^{(t)}_k \sim \text{Dirichlet}(\beta_t)$;

4) For each of the L user perspectives l, sample $\psi_l \sim \text{Dirichlet}(\eta)$;

5) For each of the $N_d$ word tokens $w_i$ in document d:
   a) sample a topic $z_i \sim \text{Multinomial}(\theta^{(d)}_d)$;
   b) sample a word $w_i \sim \text{Multinomial}(\phi^{(w)}_{zi})$;

6) For each of the T tags t in the collection D, sample $\lambda_t \sim \text{Beta}(\gamma)$;

7) For each of the $M_d$ tag tokens $t_j$ in document d created by user u;
   a) sample a flag $X \sim \text{Binomial}(\lambda_{tg})$;
   b) if ($X = 1$):
      i) Sample a topic $z_{tj} \sim \text{Uniform}(z_{m_1}, \ldots, z_{w_{d'}})$;
      ii) Sample a tag $t_j \sim \text{Multinomial}(\phi^{(t)}_{zj})$;
   c) if ($X = 0$):
      i) Sample a perspective $p_j \sim \text{Multinomial}(\theta_u)$;
      ii) Sample a tag $t_j \sim \text{Multinomial}(\psi_{pj})$;

7.2.2 Parameter Estimation

The Topic-Perspective has six parameters for estimation: (1) the document-topic distribution $\theta^{(d)}$, (2) the topic-word distribution $\phi^{(w)}$, (3) the topic-tag distribution $\phi^{(t)}$, (4) the user-perspective distribution $\psi^{(u)}$, (5) the perspective-tag distribution $\psi$, (6) and the binomial distribution $\lambda$. Several methods have been developed for estimating the latent parameters in LDA model, such as the variational expectation maximization (Blei and Jordan 2003), expectation propagation (Minka and Lafferty 2002), and Gibbs sampling.
Compared to the other two methods which are very computationally expensive, Gibbs sampling often yields relatively simple algorithms for approximate inference in high-dimensional models such as LDA. Therefore we select this approach for parameter estimation. In the Gibbs Sampling process, a Markov chain is constructed and converges to the posterior distribution on topic \( z \). The transition between successive states in the Markov chain is modeled by repeatedly drawing a topic for each observed word from its conditional probability. For the Topic-Perspective model, during the Gibbs Sampling procedure an additional Markov chain is introduced for simulating the tag generation. Inspired by the Gibbs Sampling equation for standard LDA model, we derive the sampling equations for our model. The major notations used in the following equations are explained in Table 7.1.

- Sampling equation of the word topic variables for each content word \( w_i \). (The same as standard LDA model):

  \[
P(z_{i} = k \mid w_{i} = v, z_{-i}, w_{-i}, \alpha_d, \beta_w) \propto \frac{C_{kd,-i}^{KD} + \alpha_d}{\sum_{k} C_{kd,-i}^{KD} + K\alpha_d} \cdot \frac{C_{vk,-i}^{WK} + \beta_{w}}{\sum_{v} C_{vk,-i}^{WK} + V\beta_{w}}
  \]

- Sampling equation of the tag topic variables when the switch variable \( X = 1 \):

  \[
P(x_j = 1, z_{j}^{(t)} = \tilde{z} \mid t_j = q, z_{-j}, t_{-j}, \beta_w, \beta_l, \gamma) \propto \frac{\tilde{n}_{q,-j} + \gamma}{N_{w_j}} \cdot \frac{C_{q,-j}^{KD} + \beta_l}{\sum_{q} C_{q,-j}^{KD} + \gamma T \beta_l}
  \]

- Sampling equation of the tag perspective variables when the switch variable \( X = 0 \):

  \[
P(x_j = 0, p_j = l \mid t_j = q, p_{-j}, t_{-j}, \alpha_u, \beta_l, \gamma) \propto \frac{n_{q,-j} + \gamma}{n_{q} + \tilde{n}_{q,-j} + 2\gamma} \cdot \frac{C_{u,-j}^{LU} + \alpha_u}{\sum_{q} C_{u,-j}^{LU} + L\alpha_u} \cdot \frac{C_{q,-j}^{TL} + \beta_{l}}{\sum_{q} C_{q,-j}^{TL} + \gamma T \beta_{l}}
  \]
After a set of sampling processes based on the posterior distributions calculated with the above equations, we can estimate all the parameters using the following equations:

\[
\theta_{d,k}^{(d)} = \frac{C_{kd}^{KD} + \alpha_d}{\sum_{d'} C_{kd'}^{KD} + K\alpha_d} \quad \theta_{u,l}^{(u)} = \frac{C_{lu}^{LU} + \alpha_u}{\sum_{l'} C_{lu'}^{LU} + L\alpha_u}
\]

\[
\phi_{ik}^{(w)} = \frac{C_{ik}^{WZ} + \beta_v}{\sum_{v'} C_{ik'}^{WZ} + V\beta_v} \quad \phi_{q,k}^{(t)} = \frac{C_{q,k}^{TK} + \beta_t}{\sum_{q'} C_{q,k'}^{TK} + T\beta_t}
\]

\[
\psi_{q,l} = \frac{C_{q,l}^{TL} + \beta_t}{\sum_{t'} C_{q,l'}^{TL} + T\beta_t}
\]

\[
\lambda_q = \frac{\tilde{n}_{q} + \gamma}{n_q + \tilde{n}_{q} + 2\gamma}
\]

**Table 7.1 Notations of Parameters**

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>(d, u, v, q, k, l)</td>
<td>the instance of a variable: (d) for document, (u) for user, (v) for word, (q) for tag, (k) for topic, (l) for perspective</td>
</tr>
<tr>
<td>(D, U, W, T)</td>
<td>total number of documents, users, words, and tags in the dataset.</td>
</tr>
<tr>
<td>(K, L)</td>
<td>the selected number of topics and perspectives.</td>
</tr>
<tr>
<td>(N_d, M_d)</td>
<td>the number of word tokens and tag tokens contained in document (d)</td>
</tr>
<tr>
<td>(C_{kd}^{KD})</td>
<td>the number of times that topic (k) has occurred in document (d), except the current instance</td>
</tr>
<tr>
<td>(C_{ik}^{WZ})</td>
<td>the number of times word (v) is assigned to topic (k), without counting the current instance.</td>
</tr>
<tr>
<td>(C_{q,k}^{TK})</td>
<td>the number of times tag (q) is generated from topic (k), without counting the current instance.</td>
</tr>
<tr>
<td>(C_{lu}^{LU})</td>
<td>the number of times that perspective (l) is adopted by user (u), except the current instance.</td>
</tr>
<tr>
<td>(C_{q,l}^{TL})</td>
<td>the number of times tag (q) is generated from perspective (l), without counting the current instance.</td>
</tr>
<tr>
<td>(\tilde{n}_{q})</td>
<td>the number of times that tag (q) is generated from topics ((x_q=1)), except current assignment;</td>
</tr>
<tr>
<td>(n_{q})</td>
<td>the number of times that tag (q) is generated from perspectives ((x_q=0)), except current assignment;</td>
</tr>
<tr>
<td>(\theta^{(d)})</td>
<td>a (D \times K) matrix indicating document-topic distribution.</td>
</tr>
<tr>
<td>(\phi^{(w)})</td>
<td>a (K \times W) matrix indicating topic-word distribution</td>
</tr>
<tr>
<td>(\phi^{(t)})</td>
<td>a (K \times T) matrix indicating topic-tag distribution</td>
</tr>
<tr>
<td>(\theta^{(u)})</td>
<td>a (U \times L) matrix indicating user-perspective distribution</td>
</tr>
<tr>
<td>(\psi)</td>
<td>a (L \times T) matrix indicating perspective-tag distribution</td>
</tr>
<tr>
<td>(\lambda)</td>
<td>a vector indicating the probability that each tag is generated from topics.</td>
</tr>
<tr>
<td>(\alpha_d, \alpha_u, \beta_v, \beta_t)</td>
<td>hyperparameters and priors of Dirichlet distributions.</td>
</tr>
</tbody>
</table>
7.3 Experiments

7.3.1 Topic Models for Comparison

In this section, we investigate the performance of the proposed Topic-Perspective LDA (TP-LDA) model based on a social bookmarking dataset crawled from del.icio.us. We also compare our model with two other LDA-based generative models for social annotation: the CI-LDA model and CorrLDA model.

The CI-LDA model has been used to model the generation of words and entities in (Newman et al. 2006) and the generation of words and document links in (Erosheva et al. 2004). It was adapted by Ramage et al. to jointly models the generation of word and tags for clustering purpose (Ramage et al. 2009). Figure 7.2 (a) shows the graphical representation of this model. We can see that in CI-LDA model, the tag is generated from the same source as the word: the topic of the document. Users’ impact on the generation of tags is not considered in this model.

The correspondence LDA (CorrLDA) model was originally proposed by Newman et al. (2006). It was applied to model the social tagging process by Bundschus et al. (2009). The model is graphically represented in Figure 7.2 (b). The CorrLDA model first generates word topics for a document. Then the topics associated with the words in the document are used to generate tags. Compared to the CI-LDA model, the CorrLDA model can force a greater degree of correspondence between the two information sources (in this case, words and tags). But like CI-LDA, the user information is missed in the tag generation process.

The author chooses these two models for comparison, because like the proposed Topic-Perspective model, they do the topical analysis of words and tags simultaneously.
Actually, the Topic-Perspective model is built on the CorrLDA. It extends the CorrLDA by incorporating the user factors in the tag generation process. For details about the Gibbs sampling process and equations of these two models, readers can refer to (Newman et al. 2006), where they are used for modeling the topics of words and entities in news articles.

7.3.2 Datasets

The dataset used for experiment is from a social tagging dataset collected from the Delicious website. For each web page, the full textual content was also crawled and extracted from the Internet. The web pages with no tags or containing less than 20 words were filtered out. The final dataset used for experimentation which contains 41190 documents, 4414 users, 28740 unique tags, and 129908 unique words.

7.3.3 Evaluation Criterion

The perplexity is used as the criterion for model evaluation. Perplexity is a standard measure for evaluating the generalization performance of a probabilistic model. The value of perplexity reflects the ability of a model to predict unseen data. Specifically, in our case, perplexity reflects the ability of a model to predict tags for new unseen documents. The perplexity is monotonically decreasing in the likelihood of the test data, and is algebraically equivalent to the inverse of the geometric mean of per-word (per-tag in our case) likelihood (Blei et al. 2003). A lower perplexity score indicates better generalization performance. Formally, the perplexity for a test set of $D_{test}$ documents is calculated as follows:
perplexity \(D_{test}\) = \(\exp\left\{ -\sum_{d=1}^{D_{test}} \frac{\log(p(t_d))}{\sum_{d=1}^{D_{test}} M_d} \right\} \)

\[ p(t_d) = p(x_d = 1) \sum_{k=1}^{K} p(t_d \mid z_k) p_{test}(z_k \mid d) + p(x_d = 0) \sum_{l=1}^{L} p(t_d \mid p_l) p_{test}(p_l \mid u) \]

In the above equation \(t_d\) is a tag included in the test document \(d\). The probabilities \(p(t_d \mid z_k)\), \(p(t_d \mid p_l)\), and \(p(x)\) are learned from the training process, and \(p_{test}(z_k \mid d)\) and \(p_{test}(p_l \mid u)\) are estimated through a Gibbs Sampling process on the test data based on the parameters \(\phi^{(w)}, \phi^{(t)}, \psi, \) and \(\lambda\) learned from training data.

### 7.3.4 Experimental Setup

The Topic-Perspective model has six Dirichlet prior parameters. The author tested a serial of values for each parameter and found that their effect on the perplexity value is little. Previous research also found that these parameters only affect the convergence of Gibbs sampling but not much the output results (Ramage et al. 2009, Zhou et al. 2008). So the author set \(\alpha_d=0.3, \alpha_u=0.3, \beta_w=0.05, \beta_t=0.05, \eta=0.05, \gamma=0.5\) for all experiments.

The remaining question is how to select the number of topics \(K\) and the number of perspectives \(L\). The author first fixed the number of perspectives to a certain number, and then tested the perplexity of the trained model on the test data for different topic numbers. The smallest topic number which leads to the minimum or near minimum perplexity is selected. After the topic is chosen, the perspective number is selected similarly based on the perplexity. Figure 7.3 shows a plot of perplexities on five different settings of \(K\), when the perspective number is fixed to 80. We can see that in general the perplexity scores for all topic number settings decrease along the iterations. The algorithm tends to converge after about 40 iterations. Along the iterations, larger setting of topic number
always leads to smaller perplexity value from the start, and indicating a better prediction performance. But the effect of increase in topic number on perplexity value gets smaller when the topic number gets larger. When the topic number set to 160, the perplexity value actually goes up. Therefore, the topic number $K$ is set to be 80 which leads to the minimum perplexity among the five settings.

Figure 7.3  The perplexities over the iterations for five settings of topic number when perspective number $L=80$

Figure 7.4  The perplexities over the iterations for five settings of perspective number when topic number $K=80$
The situation for selecting perspective number is similar. Figure 7.4 displays the plot of perplexities for five settings of perspective number $L$ when topic number is set to 80. Still the perspective number $L=80$ leads to the minimum perplexity score. And when $L$ increases to 160, the perplexity value sharply goes up. So for the final experiment, both topic number and perspective number are set to 80.

7.4 Results

7.4.1 Tag Perplexity

The author compares the tag prediction abilities of the Topic-Perspective model with CorrLDA model and CI-LDA model based on the perplexity value. Figure 7.5 plots the perplexity results for each model over different topic numbers. The perspective number of TP-LDA was set to be 80, and the iteration numbers for all three models were set to 80. We can see that, before $K=80$ TP-LDA constantly performs better than other two models especially when the topic number is small. For all three models, larger topic number generally leads to smaller perplexity scores. This is because the increased topic number reduces the uncertainty in training. However, the effect of topic number on the three models’ performance is different. TP-LDA model is least affected by the topic number. Especially, when the topic number increases to 160, its perplexity value grows up. This is because TP-LDA incorporates the users’ perspective information into the tag generation process, and the predicted tags do not completely account on document topics.

From figure 7.5, we can also see that CorrLDA performs worse than CI-LDA. This
is because CI-LDA uses both content words and tags to learn document topics, but CorrLDA only learns topics from content words. Recall that, in CorrLDA, only the topics learned from content words are used to generate tags. Therefore, in CI-LDA model, the topics learned from the training data are more associated with tags and thus are more effective for tag prediction. This result further indicates the difference of words and tags in topical structure. Although CI-LDA generates better results, experimental results show that CI-LDA’s word topics and tag topics are too decoupled. Little correspondence can be found between the words and tags generated to represent the same topic. The TP-model overcomes this limitation without sacrificing the performance.

![Figure 7.5](image.png)

Figure 7.5 The perplexity results of CorrLDA, CI-LDA and TP-LDA (Topic-Perspective Model) for topic number \( K = 10, 20, 40, 80, 160 \)

7.4.2 Discovered Topics and Perspectives

The topics and perspectives discovered by the Topic-Perspective model can be analyzed by examining the top words and tags assigned to each topic and the top tags assigned to each perspective.
Table 7.2  A subset of discovered topics

<table>
<thead>
<tr>
<th>Topic ID</th>
<th>Top words</th>
<th>Top tags</th>
</tr>
</thead>
<tbody>
<tr>
<td>7</td>
<td>war world militaries nation state force govern unite iraq countries international israel american armies peace</td>
<td>politics history world international war military activism poverty information africa islam government middleeast europe humanright</td>
</tr>
<tr>
<td>13</td>
<td>mountain fish camp boat adventure sea river park trail climb ski new lake gear sail</td>
<td>travel camp backpack hike photography climb sail photo knot boat gear nature adventure ski kayak</td>
</tr>
<tr>
<td>15</td>
<td>movie film star video dvd man episode new release trailer love review girl fan season</td>
<td>movy video entertainment film music review movie humor television medium fun funny cinema stream comicstrip</td>
</tr>
<tr>
<td>16</td>
<td>church god christian beer bible jesus religion faith new catholic christ life religion holi john</td>
<td>religion bible christianity christian church history buddhism mythology atheism theology spirituality philosophy apologetics catholic culture</td>
</tr>
<tr>
<td>20</td>
<td>window linux software file microsoft install mac computer user server program your run desktop disk</td>
<td>software window linux mac freeware osx utility ubuntu apple backup download sysadmin virtualization security opensource</td>
</tr>
<tr>
<td>25</td>
<td>law legal copyright inform public patent right state court govern lawyer act protect file agency</td>
<td>law copyright legal government internet privacy patent technology security politics research right p2p tech plagiarism</td>
</tr>
<tr>
<td>28</td>
<td>recipe food cook cup coffee chocolate cake eat tea cheese bake add bread make water</td>
<td>food recipe cook howto health drink coffee nutrition vegetarian collection restaurant tea bake kitchen diet</td>
</tr>
<tr>
<td>46</td>
<td>university student school education studies college science teacher program teach course institution graduate academ department</td>
<td>education teach learn school research science elearn university resource academic college kid study math lessonplan</td>
</tr>
<tr>
<td>49</td>
<td>book write author stories publish writer read fiction comic chapter novel poetries edit amazon poem</td>
<td>book write ebook literature comicstrip read publish library comic poetry tutorial webcomic selfpublish author scifi fiction</td>
</tr>
</tbody>
</table>

Despite of the lack of quantitative assertions, we observe generally high semantic correlations among the top words and tags for the each topic, and high correspondence between the words and tags for the same topic. The themes of the discovered topics are diverse, and mostly related to the hot subjects, such as web design, programming, traveling, shopping, education, politics, etc. Table 7.2 displays the top ten words and tags of a random subset of discovered topics. Because of the coherent semantics of the words
and tags for each topic, the theme of each topic is obvious. For instance, Topic 7 is about the war and politics, Topic 13 is on outdoor activities, Topic 15 is associated with movies, and so on.

Table 7.3   A subset of discovered perspectives

<table>
<thead>
<tr>
<th>Perspective ID</th>
<th>Top tags</th>
</tr>
</thead>
<tbody>
<tr>
<td>11</td>
<td>reference guide multimedia list help codec portal emulator comparison boot upload anonymous virtual organize proxy</td>
</tr>
<tr>
<td>24</td>
<td>competition switch event blogroll likeddesign inspire wysiwyg creative artistresource ria tagthese domainname affiliate editorial cooky</td>
</tr>
<tr>
<td>32</td>
<td>conference metadata openacce tag association folksonomy censorship preservation digitallibrary sheetmusic librarian secondlife rfid directory digitalgame</td>
</tr>
<tr>
<td>36</td>
<td>search link portal directory list system indie tag rock about customize current label usenet ezine kaizen synchronization</td>
</tr>
<tr>
<td>47</td>
<td>bookmark readlate quickd engl401 ircbot meetup shirt favoritesmenu emergent nikon twincity tattoo punk oreilly jobsite simplicity</td>
</tr>
<tr>
<td>51</td>
<td>good publication product stuff thesis awesome mypublication gobacktothis florida giztag sidebar nicedesign travelinfo longdistancetrip sourdough myprofile</td>
</tr>
<tr>
<td>52</td>
<td>developer article example onlinekit issuetrack fstUDIO aggregation swifteam webbuild backend asus wtf guideline communication swiftmobile</td>
</tr>
<tr>
<td>58</td>
<td>compute archive multimedia wireless app macintosh directory communication datum codec freedom admin cheat classic mystuff list</td>
</tr>
<tr>
<td>64</td>
<td>toread todo totry todownload webdevelope mind tobrowse tobuy tocheck conference landscape frequentlyuse epge usefulsoftware intelligence</td>
</tr>
</tbody>
</table>

The Topic-Perspective model also discovers the user perspectives from the tags. The perspectives are more complicated than topics. The correlation among the tags assigned to each perspective is not as obvious as those for topics. This is because, unlike topic, a user perspective does not reflect a pure aspect of tags. Each perspective may combines several user factors of social tagging, such as user’ domain background, preference, interest, motivation, etc. Despite of the complexity of perspectives, we can still identify
some patterns by examining the tags assigned to each perspective. Table 7.3 lists a subset of discovered perspectives and their top tags. We can see that the tags assigned to perspectives are very different from those assigned to topics. If we look back to Bischoff’s classification of tags in Table 2.1, it is apparent that the tags assigned to perspectives generally belong to the categories other than Topic. For instance the tags for Perspective 11 are mostly for describing the documents’ type, tags in Perspective 51 are used for opinion-expression and self-reference, and tags for Perspective 64 are used for task organization and self-reference.

7.4.3 The Generation Sources of Tags

The Topic-Perspective model use an additional variable $\lambda$ ($0 < \lambda < 1$) to record the probability that each tag is generated from topics or user perspectives. Greater value of $\lambda$ indicates a higher probability that the tag is generated from document topics and vice versa. Table 7.4 lists some example tags with $\lambda = 1$, 0.5 and 0.

<table>
<thead>
<tr>
<th>$\lambda=1$</th>
<th>$\lambda=0.5$</th>
<th>$\lambda=0$</th>
</tr>
</thead>
<tbody>
<tr>
<td>library shop internet</td>
<td>palmpre audiomagazine mathware</td>
<td>tag app interest archive</td>
</tr>
<tr>
<td>research socialnetwork</td>
<td>postapocalyptic educause vomit</td>
<td>toread datum code todo</td>
</tr>
<tr>
<td>statistic ruby ajax</td>
<td>nwiqpartn singlespe masterproof</td>
<td>webservice directory list</td>
</tr>
<tr>
<td>javascript webdev culture</td>
<td>richmullin sundial selenium showstep</td>
<td>guide link portal training</td>
</tr>
<tr>
<td>music health graphic math</td>
<td>webmath randynewman immortalism</td>
<td>site track article</td>
</tr>
<tr>
<td>security firefox cs politics</td>
<td>malazan architecturalproduct</td>
<td>reference web20 online</td>
</tr>
<tr>
<td>recipe photography</td>
<td>fotologserevista biblioteque caribbean</td>
<td>search tool free cool</td>
</tr>
<tr>
<td>europeana</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
The tags with $\lambda=1$ are completely generated from topics and not affected by users’ perspectives. We can see that these tags can clearly and objectively reflect the topics of the annotated documents. Contrarily, the tags with $\lambda=0$ are totally generated by users’ perspectives. It is clear to see that these tags contribute little to reflecting the topics of the annotated documents. They are created by users for other purposes other than identifying topics. An interesting observation was made on the tags with $\lambda=0.5$, which are equally influenced by document topics and user perspectives. From table 5, we can see that these tags are mostly terms invented by the users which cannot be found in dictionaries. Most of them are phrases with no space between words, such as “audiomagazine”. Different from tags with $\lambda=0$, these tags are actually related to the document topics. In other words, they are created to describe the topics in a personal way.

7.5 Future Applications

The results of the proposed model can be applied for tag prediction. Given a new document, based on the parameters estimated by the model, the tags can be predicted in two ways. For general tag prediction, we can only consider the tags with high topic probability and filter out tags with high perspective probability. The likelihood of a tag $t$ for a test document $d$ is:

$$p(t \mid d) = \sum_{k=1}^{K} p(t \mid z_k) p(z_k \mid d)$$

$p(t_d \mid z_k)$ is given by the topic-tag distribution, and $p(z_k \mid d)$ is estimated online based on the parameters learned from the training process. For personalized tag prediction, namely if we want to predict the tags that could be created for a document by a specific user $u$
who has also appeared in the training dataset, we can calculate the likelihood of a tag \( t \) for a test document \( d \) as follows:

\[
p(t \mid d, u) = \lambda p(t \mid z_k) p(z_k \mid d) + (1 - \lambda) \sum_{l=1}^{L} p(t \mid p_l) p(p_l \mid u)
\]

, in which \( p(t \mid z_k), p(t \mid p_l), p(p_l \mid u) \) and \( \lambda \) are learned from the training process, while \( p(z_k \mid d) \) are estimated online.

The parameters learned from the model to can also be used to enhance the performance of information retrieval (IR). For general information retrieval, we can enhance the IR language model with the tags of high topical probability and the topical structures of the words and tags. For personalized information retrieval we can further expand the IR language model with tags of high perspective probability and the users’ perspectives.

### 7.6 Conclusion

This chapter proposes a Topic-Perspective LDA model to simulate the tag generation process. By modeling the tag generation and word generation process separately and incorporating the user information into the tag generation process, the proposed model is able to model the social annotation system in a more meaningful way and achieve better generalization performance than other models. Besides, this model also generates useful information about the topical structures of tags and words, as well as the influences of document topics and user perspectives on different tags. The results derived from this model can be utilized for automatic tag recommendation, personalized web search and other web mining applications.
7.7 Research Question Tested

*Question 4: How to build a topic model for the tagged web?*

This chapter proposes a novel probabilistic generative model, called Topic-Perspective Model, to model the tagged web. The Topic-Perspective model reflects the real social tagging process and represents all related entities (users, documents, words, and tags) and latent variables (topics, user perspectives) in a unified model.
CHAPTER 8: CONCLUSIONS

The research of this thesis is dedicated to investigating the role that social tagging data can play in enhancing the performance of web mining and search applications. The research is inspired by the observed properties of social tagging that make it potentially valuable for web mining and search applications. These properties include its indexing efficiency, its adaptability to emerging concepts and vocabularies, its consistency between indexers and searchers, and its rich information about users. At the same time, the author also realizes the factors that make it challenging to utilize social tagging data for web mining and search. These factors include the quality issue of social tagging stemming from its uncontrolled nature, the prevalence of personal and subjective tags, as well as the tripartite relation embedded in the social tagging network. The research of this thesis revolves around exploiting the valuable properties of social tagging for web mining and search and overcoming the challenges during the research process.

8.1 Social Tags as Index Terms

One of the research goals of this thesis is to investigate whether social tags are effective document features which can be used to represent and index web documents in web mining and search applications. To attain this research goal, the author compares social tags with other type of index terms, including expert-created subject terms, author-created keywords and description, as well as the content words of web pages. The key findings are summarized as follows.

First, it turns out that user tags can cover almost half of the subject terms created by
experts, indicating that users are capable of creating some high-quality index terms. But users’ tag vocabulary is much more diverse than expert-created subject terms. Most of user tags cannot be found in the subject terms. The tags that do not overlap with subject terms represent the divergence between the terms used by users and experts when they describing web documents. This indicates that user tags could be used to expand the access points of resources already cataloged and indexed by experts.

Compared to expert-created subject terms, author-provided keywords seem to be more close to user tags in nature. In our study, about one third of the tags are covered by author keywords. Although user tags are very diverse compared to expert-created subject terms, they bear higher convergence than author-generated keywords. For the same collection of web documents, the number of unique keywords is about five times of the number of unique tags. This may be because of the stable pattern that has been identified in social tagging (see section 2.1.2). Namely, after a webpage has been tagged by a certain number of users, a consensus about the terms used for describing the webpage is formed. This nature makes social tags more manageable and favorable than author keywords as index terms.

Social tags and author keywords are also compared with content words. It turns out that that both of them can add new information to existing page content. The value of the additional information provided by keywords and social tags is further demonstrated through a clustering study, which proves that both tags and author keywords can be used to significantly improve content-based clustering. Besides, tags are more effective than author-provided metadata when used as independent document features from document content. Although only clustering method is used for experimentation, the conclusions
can be extended to other web mining and search applications.

8.2 Exploiting the Tripartite Network of Social Tagging

A social tagging system can be modeled with a tripartite network of users, documents and tags. The tripartite network provides rich information to learn document topics, user interests and tag semantics. However, how to effectively model and utilize the tripartite network is challenging. In this thesis, the author proposes two approaches to model the tripartite network of social tagging.

The first approach represents the social tagging network as a tripartite graph, with three types of vertices (user, document, tag) and three types of undirected binary edges (user-document, user-tag, and document-tag) (See equation (5.1)). Based on this model, the author proposes a novel clustering method called Tripartite Clustering, which clusters documents, users and tags simultaneously based on the three types of binary linkages in the graph. The Tripartite Clustering algorithm significantly outperforms the content-based K-means. The author also compares the Tripartite Clustering method with an enhanced K-means algorithm which incorporates tags and users as document features for clustering, and a Link K-means algorithm which utilizes both content words and link information extracted from the social tagging network for document clustering. Experimental results show that Tripartite Clustering method significantly outperforms Link K-means and achieves comparable results to the enhanced K-means approach.

In the second modeling approach, the tripartite social tagging network is represented with a 3-uniform hyper-graph which uses a hyper edge to represent the ternary relation among users, documents and tags (See equation (6.1)). In this way, this model avoids projecting the tripartite relation into three binary relations, and thus provides a more
integral approach to represent the tripartite social tagging network. Based on this hypergraph model, the author propose a personalized search framework (TripleQE) which fully relies on the tripartite relationship among users, tags and web documents for personalized search. Experiments demonstrate that the proposed personalized search framework is more effective than traditional language model based search methods. Moreover, when combined with language model based search methods, the proposed search framework can produce better search results.

### 8.3 Topic Models for Social Tagging System

In the text mining and information retrieval community, topic modeling is becoming popular as an approach for discovering topical structures from documents. It is based on the development of probabilistic generative models which simulates the generation of document content. In this thesis, the author applies topic models to simulate the generation of social tags and accordingly discover the topical structures of documents and users’ tagging perspectives. Based on the nature of social tagging process, the author proposes a Topic-Perspective LDA model which models the generation of tags and words in a unified model and incorporates the user variable into the tag generation process. Experimental analysis shows that this comprehensive model has better generative ability than topic models proposed in existing literature. Besides, this model also generates more useful information about document topics, user perspectives, and the impact of document topics and user perspectives on tag generation, which is useful for distinguishing topical tags from personal and subjective tags.
8.4 Contributions

Firstly, this thesis sheds light on the value and obstacles of using social tags to represent, index, organize and retrieve web documents. The empirical studies conducted by the author provide some tangible evidence for those who are concerned with the feasibility and effectiveness of social tagging as an indexing approach.

Secondly, the author developed several approaches to demonstrate how the social tags can be exploited for web mining applications, such as web clustering. The proposed novel Tripartite Clustering method shows how the tripartite network of social tagging can be utilized to enhance the clustering performance without using document content information. It also exemplifies a new clustering approach which clusters more than one object all at once based on the link information.

Thirdly, the author proposed a novel personalized search framework based on a hyper-graph model of social tagging system. The proposed framework demonstrates an efficient and effective approach of utilizing the user information learned from the social tagging system for personalized search.

Finally, the author designed an advanced topic model for the social tagging system. The so-called Topic-Perspective model not only reflects the real tag generation process, but also integrates different variables involved in the social tagging process into a unified model. It reflects the interactions among different variables in a social tagging system. It can be used to identify the topical structures of tagged documents and the perspectives of taggers. Moreover, it can be used to estimate whether the generation of a tag is more influenced by document topics or taggers’ personal factors.
LIST OF REFERENCES


Qi X. and Davison B.D., “Knowing a web page by the company it keeps” in Proc.15th ACM Int. Conf. Information and Knowledge Management, New York, NY, 2006, pp. 228 - 237.


Yoo I., Hu X., and Song I.-Y., “Integration of semantic-based bipartite graph representation and mutual refinement strategy for biomedical literature clustering” in


VITA

CONTACT
Caimei Lu, caimei.lu@gmail.com, 215-900-4164

EDUCATION
Ph.D. in Information Science
College of Information Science and Technology, Drexel University 2006-2011

Master in Information Science
Department of Information Management, Nankai University, China 2003-2006

Bachelor in Information Management Systems
School of Management, Lanzhou University, China 1999-2003

RESEARCH INTERESTS
Text Mining, Information Retrieval, and Social Network Analysis, with particular interests in developing models and systems that utilize social media data to enhance resource discovery and web search.

RESEARCH EXPERIENCE
Research Assistant, College of Information Science and Technology, Drexel University 2006-2011

TEACHING EXPERIENCE
Lecturer, Info210 Database Management Systems
College of Information Science and Technology, Drexel University 2010 Fall

Teaching Assistant, Info210 Database Management Systems
College of Information Science and Technology, Drexel University 2010 Summer

SELECTED PUBLICATIONS