CHARACTERIZING AND EVALUATING USERS' INFORMATION SEEKING

BEHAVIOR IN SOCIAL TAGGING SYSTEMS

by

Tingting Jiang

B.S., Wuhan University, 2003

M.S., Wuhan University, 2005

Submitted to the Graduate Faculty of

School of Information Sciences in partial fulfillment

of the requirements for the degree of

Doctor of Philosophy

University of Pittsburgh

2010

UNIVERSITY OF PITTSBURGH

SCHOOL OF INFORMATION SCIENCES

This dissertation was presented

by

Tingting Jiang

It was defended on

December 8, 2010

and approved by

Committee Chair: Daqing He, Ph.D., Associate Professor, School of Information Sciences,

University of Pittsburgh

Leanne Bowler, Ph.D., Assistant Professor, School of Information Sciences, University of Pittsburgh

Michael Lewis, Ph.D., Professor, School of Information Sciences, University of Pittsburgh

Jin Zhang, Ph.D., Professor, School of Information Studies, University of Wisconsin - Milwaukee

Dissertation Advisor: Sherry Koshman, Ph.D., Assistant Professor, School of Information Sciences,

University of Pittsburgh

CHARACTERIZING AND EVALUATING USERS'

INFORMATION SEEKING BEHAVIOR IN SOCIAL TAGGING SYSTEMS

Tingting Jiang, PhD

University of Pittsburgh, 2010

Social tagging systems in the Web 2.0 era present an innovative information seeking environment succeeding the library and traditional Web. The primary goals of this study were to, in this particular context: (1) identify the general information seeking strategies adopted by users and determine their effectiveness; (2) reveals the characteristics of the users who prefer different strategies; and (3) identify the specific traits of users' information seeking paths and understand factors shaping them. A representative social tagging system, Douban (http://www.douban.com/) was chosen as the research setting in order to generate empirical findings.

Based on the mixed methods research design, this study consists of a quantitative phase and a qualitative phase. The former firstly involved a clickstream data analysis of 20 million clickstream records requested from Douban at the footprint, movement, and track levels. Limited to studying physical behavior, it was complemented by an online survey which captured Douban users' background information from various aspects. In the subsequent qualitative phase, a focus group gathered a number of experienced Douban users to help interpret the quantitative results.

Major findings of this study show that: (1) the general strategies include

encountering, browsing by resource, browsing by tag, browsing by user/group, searching, and monitoring by user/group; (2) while browsing by resource is the most popular strategy, browsing by tag is the most effective one; (3) users preferring different strategies do not have significantly different characteristics; and (4) on users' information seeking paths these exist two resource viewing patterns – continuous and sporadic, and two resource collecting patterns – lagged and instant, and they can be attributed to user, task, and system factors.

A model was developed to illustrate the strategic and tactic layers of users' information seeking behavior in social tagging systems. It offers a deep insight into the behavioral changes brought about by this new environment as compared to the Web in general. This model can serve as the theoretical base for designing user-oriented information seeking interfaces for social tagging systems so that the general strategies and specific tactics will be accommodated efficiently.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

1.0 INTRODUCTION

1.1 BACKGROUND

Information seeking is a fundamental human process. We look for information, through interacting with manual information systems (e.g. traditional libraries) or computer-based information systems (e.g. the Web), to change our state of knowledge (Marchionini, 1995; Wilson, 2000). In either context, it is always desirable that the systems offer users abundant information resources with comprehensive coverage. However it is also important that these resources are well arranged and easily accessible in order to make the information systems usable and truly useful.

People have been acquiring books, periodicals, maps, paintings, and other materials in the libraries since the 1800s. And professional catalogers have devoted efforts to creating bibliographic records for these library collections, the principal purposes of which are to facilitate storage of and retrieval in large collections (Taylor, 2004). Formal classification systems (e.g. Dewey Decimal Classification and Library of Congress Classification) and subject heading systems (e.g. Library of Congress Subject Headings and Medical Subject Headings) have been the primary tools for organizing and providing access to physical resources, and metadata schemes (e.g. Dublin Core) have emerged specifically for describing electronic resources (Chan, 2007).

The introduction and growth of the Web have resulted in profound changes in human information seeking. Although the Web accommodates a considerable amount of structured information collections to which conventional cataloging practices are still applicable, it is on the whole a heterogeneous space with vast volumes of Web documents bearing great variability in terms of format, size, focus, and quality, etc. Furthermore, these documents are interconnected via hyperlinks and can be added, altered, and deleted anytime (Rasmussen, 2003).

In spite of the complexity of the Web, we see exceptional efforts in "cataloging the Web", such as the notable Yahoo! directory, a hierarchical subject index of websites (Callery, 1996). Browsing such a human-powered Web classification system is a common way to discover interesting resources, while most Web users still prefer to locate their needed information by means of search engines (Rasmussen, 2003). Search engines, in a certain sense, are also metadata schemes, only less structured than others (Younger, 2002). They automatically index the content of the webpages by extracting metadata from various fields or even the full texts, then match users' search keywords with the metadata to determine search results, and rank the results algorithmically to display the most relevant ones on the top of the result sets (Glossbrenner, 2001).

The most recent revolution in the information landscape, namely, Web 2.0, not only inherits the diversity and dynamics of the Web, but exhibits even greater complexity for being a participatory platform where general Web users are allowed to create, store, and share their own information collections (Marlow *et al.*, 2006). Correspondingly, users are also driven to assume the responsibilities of describing and categorizing the items in their collections and making them findable to others. Such an innovative cataloging practice is quite simple and known as "tagging" –

users adding metadata or keywords to information resources (Golder & Huberman, 2006). Tagging is basically an individual user behavior since users tag according to their personal understanding and in a distributed manner. However, the "social" aspect of tagging consists in the facts that users share tags and that tags are aggregated into a social classification system called "folksonomy", which are enabled in social tagging systems.

A "social tagging system" is a general term that refers to any Web 2.0 site dedicated to preserving information resources collected by users and basically relying on tagging to catalog such resources (Kalbach, 2007). These two features distinguish social tagging systems from other websites which also support social tagging. Amazon.com[1], for example, has introduced customer tagging to supplement the well-established "departments" of products. With the types of information resources varying, social tagging systems further divide into the following major categories[2]:

- *Social bookmarking systems*: keeping Web users' bookmarks of webpages that are formerly saved to local Web browsers. The term "social bookmarking" is sometimes used interchangeably with "social tagging" because the latter was firstly seen in Delicious[3], one of the earliest and most popular social bookmarking systems (Hammond *et al.*, 2005). Delicious opens to the general public, like many other systems, e.g. Reddit[4], Diigo[5], and StumbleUpon[6] etc. But Dogear is a

---

[1] http://www.amazon.com/
[2] http://en.wikipedia.org/wiki/Social_software
[3] http://delicious.com/
[4] http://www.reddit.com/
[5] http://www.diigo.com/
[6] http://www.stumbleupon.com/

corporate-wide system within the Intranet of IBM (Millen *et al.*, 2006).

♦ *Social citation systems*: especially targeting scientists and scholars and helping them organize the citations or references of academic publications for work or research purposes. One of leading social citation systems, CiteULike[7], has supported users to upload the PDF files of the papers or articles, while other similar services such as Cannotea[8] and Bibsonomy[9] are still limited to collecting citations.

♦ *Social library systems*: enabling people to build virtual shelves of personal collectibles which are mainly books, music records, and movies so far. The idea is for users to easily keep track of what they own or have interests in. In other words, they do not really read, listen, or watch within the systems. LibraryThing[10], Discogs[11], and IMDb[12] are representative social library systems respectively specializing in books, music, and movies.

♦ *Social guide systems*: aggregating users' recommendations for sightseeing attractions, adventure destinations, or other places in the real world, such as restaurants, hotels, coffee shops, and Wi-Fi hotspots, etc., usually accompanied by peer voting of the places recommended. Example social guide systems include Thoos[13], Socialguides[14], Tripist[15], and so forth.

♦ *Multimedia sharing systems*: allowing users to store and share the actual digital objects which they

---

[7] http://www.citeulike.org/
[8] http://www.connotea.org/
[9] http://www.bibsonomy.org/
[10] http://www.librarything.com/
[11] http://www.discogs.com/
[12] http://www.imdb.com/
[13] http://www.thoos.com/guides/
[14] http://www.socialguides.com/
[15] http://www.tripist.com/

create or collect from elsewhere, including photos, videos, podcasts, PowerPoint slides, blog posts and other multimedia files. This makes them a little different from the above systems which only contain the pointers to various objects. Flickr[16] for photo sharing and YouTube[17] for video sharing are the two globally famous multimedia sharing systems at present.

It is certainly difficult to exhaust the long list of social tagging systems. Accordingly there is a large amount of literature on this new area of research which has increased dramatically during the past five years. Trant (2009) reviewed about 180 relevant academic papers and identified three broad research foci, i.e. tagging behavior, folksonomies, and socio-technical frameworks. Tagging behavior studies are interested in why users tag and especially, how they tag, including tag usage, frequency, distribution, and co-occurrence etc. In investigating folksonomies, researchers agree that they excel taxonomies for being economic, current, flexible, and democratic, while even more attention has been drawn to their insufficiencies in terms of vocabulary and structure as well as the resulted negative impacts. The last focus, socio-technical frameworks, examines the tools provided by social tagging systems, such as navigation, ranking, visualization, social networking, and so on.

1.2 PROBLEM STATEMENT

As more and more users register with various social tagging systems, the Web is actually

---

[16] http://www.flickr.com/

[17] http://www.youtube.com/

experiencing the rapid self-growth of numerous information repositories, in many of which user-contributed content has reached a substantial amount. For example, there have been more than 150 million unique URLs bookmarked on Delicious, about 45 million books cataloged in LibraryThing, and 4 billion photos uploaded to Flickr, according to some recent statistics[18].

Given the abundance or sometimes even excess of information in social tagging systems, can users, as information seekers, find their interested resources effectively? This is a crucial question that has an immediate influence on the level of users' participation in the tagging activity, since they tag to retrieve information in their personal collections, and more importantly, to find information shared by others (Kalbach, 2007).

There is only preliminary discussion concerning the above question in the existing literature, with contradicting opinions. Proponents of social tagging think highly of folksonomies' ability to offer information discovery that leads users to unexpected yet potentially useful information (Kroski, 2005). Critics, on the contrary, have stated about folksonomies that "when it comes to findability, their inability to handle equivalence, hierarchy, and other semantic relationships causes them to fail miserably at any significant scale" (Morville, 2005, pp.139). Unfortunately, it is difficult to determine which side should be given more credit since neither of them provides any objective evidence to support their arguments.

What's more, it is obvious that the whole controversy is confined to folksonomies. In a pioneering study on folksonomy and exploratory search, Jiang and Koshman (2008) understood the

---

[18] http://en.wikipedia.org/wiki/Delicious_(website)
   http://en.wikipedia.org/wiki/LibraryThing
   http://en.wikipedia.org/wiki/Flickr

folksonomy as an information architecture which played an essential role in locating resources for vague information needs. However, they further suggested that information seeking in social tagging systems as a matter of fact was characteristic of the interplay of the four information seeking strategies – encountering, browsing, searching, and monitoring, and the three social tagging elements – resources, users, and tags. This illuminating study, nevertheless, is still conceptual.

To sum up, previous studies related to information seeking in social tagging systems demonstrate two major deficiencies. First, the research scope is narrow. Information seeking activities involving no use of folksonomies, which may be very common in today's systems, are basically neglected. And second, empirical research is rare, probably because the exploration of this topic is still at an early stage.

1.3 RESEARCH QUESTIONS

This dissertation study addresses the above problems by positioning users' information seeking behavior within a broader framework and externalizing their behavior to generate empirical findings. That is to say, real-world data from users' everyday interaction with social tagging systems as an integrated information seeking environment will be collected and analyzed, with full consideration to the complex interplay. Below are the research questions answered in this study:

◆ Research Question 1 (RQ 1):

What are the general information seeking strategies adopted by users in social tagging

systems and how effective are they in helping users find information resources of interest?

◆ Research Question 2 (RQ 2):

For each information seeking strategy identified, is it possible to generalize the characteristics of the users who prefer to adopt it? If yes, what are these characteristics?

◆ Research Question 3 (RQ 3):

What are the specific traits of users' information seeking paths in social tagging systems and what are the factors contributing to the formation of their information seeking paths?

In RQ 1, the "general information seeking strategies" specifically refer to the high-level planning of actions designed for an information need. RQ 2 is the follow-up question of RQ 1, trying to relate users' backgrounds to their most frequently adopted strategies. The term "information seeking path" in RQ 3 can be understood as a sequence of actions taken by the user within a social tagging system in order to satisfy certain information needs. The notion of path emphasizes the dynamics of user behavior, with each action on a path being the cause of the next one until the path ends.

1.4 SIGNIFICANCE OF THE STUDY

The Web is an ever-changing world where technological developments constantly introduce new ways of organizing and accessing information. While Web users still depend on search engines for information seeking to a great extent, they are meanwhile attracted to an increasing variety of

information systems with novel characteristics. They will make behavioral adaptions correspondingly to get themselves accustomed to the new mechanisms so that they can make better use of the new capabilities.

This dissertation study is one of the first to investigate information seeking behavior in social tagging systems which represent the most recent tide of innovation on the Web. Its significance not only lies in the attention given to such under-researched yet important context, but more essentially in that it deals with the apparent topical and methodological limitations in current information seeking literature.

On the one hand, this study presents a holistic perspective that takes the whole information seeking process into account. Despite the multidimensional nature of human information seeking, individual studies in the past usually fail to explore how people act, think, and feel in one general research schema (Martzoukou, 2005). This study, in contrast, probes into user behavior and at the same time tries to reveal various user characteristics with an impact on the behavior. It further gives full consideration to the properties of social tagging systems as a more diverse and dynamic information seeking environment by looking beyond searching behavior which has been the exclusive focus of most former research. All of the four major information seeking strategies – encountering, browsing, searching, and monitoring (Wilson, 1997; Bates, 2002) – are regarded here to create a comprehensive picture of the behavioral traits of social tagging system users.

On the other hand, this study contributes to the development of research methods for the area of information seeking. Because of a prevalent focus on information searching, most existing methods correspondingly apply exclusively to searchers and their searching behavior. The variables

examined basically include terms, queries, and searching episodes that are specific to the interaction between users and search engines (Jansen, 2006). This study is instead concerned with users' information seeking paths and proposes to analyze them at footprint, movement, and track levels, which will be specified in the methodology chapter. Not being restricted to any particular information seeking strategy, such an analysis framework has wide applicability to different information systems including social tagging systems in investigating how users approach their information goals step by step in virtue of single strategies or the combinations of multiple strategies.

2.0 LITERATURE REVIEW

Since this dissertation study involves information seeking behavior (ISB) as the topic and social tagging systems as the context, the following literature review naturally divides into these two main parts. ISB research, which has a long history, has produced a large quantity of publications. The most relevant ones to this study are included here. They are organized into two sections. Section 2.1 concentrates on the theoretical foundations of the area, especially expounding strategical ISB theories that are most relevant to this study. Then Section 2.2 provides an overview of the existing empirical studies in the area, with an emphasis on the Web environment and searching as the most popular way of information seeking on the Web.

Research on social tagging systems nevertheless has just started in recent years when Web 2.0 became an important phenomenon. Although different systems have been examined in terms of users' tagging behavior, the folksonomy, or socio-technical frameworks, this study is more interested in scholarly publications that treat of how users look for information or any tool facilitating their information seeking activities. After an introduction to the characteristics of Web 2.0 as the background in Section 2.3.1, Section 2.3.2 decomposes the social tagging system into the three basic components: resources, users, and tags.

2.1 INFORMATION SEEKING BEHAVIOR: THEORETICAL FOUNDATIONS

2.1.1 Concepts

According to a co-citation analysis of all the articles published in 21 Library and Information Science (LIS) journals between 1991 and 2004, information seeking and retrieval (ISR) is one of the two continuously dominating research areas, the other being informatics (Astrom, 2007). The ISR literature, as found by the same analysis, demonstrates a shift of focus from the system side to the human side, and user behavior in the broader information seeking and use process has become a fundamental theme of research.

Information seeking behavior is among the most frequently seen terms in the ISR literature, yet used at inconsistent levels in different contexts. Given its significance to this dissertation study, it will be helpful to first of all clarify the scope of ISB and to distinguish it from other closely related concepts. Below are two widely recognized definitions of ISB.

- Marchionini (1995): information seeking is "a process in which humans purposefully engage in order to change their state of knowledge" (p. 5-6).

- Wilson (2000): "information seeking behavior is the purposive seeking for information as a consequence of a need to satisfy some goal" (p. 49).

These definitions agree that ISB is triggered by information needs. Wilson (1981) pointed out that the needs for information in turn stemmed from the lack of self-sufficiency in our everyday

life and work. We may know where exactly the gaps exist in our current knowledge or not, which means that information needs can be conscious or unconscious. ISB is all about gathering relevant information for a certain need, but whether the need can be satisfied also depends on how the information is used (Devadason & Lingam, 1996). Information use behavior, in addition to physical actions such as taking notes or marking texts, may also take the form of thinking, comparing, deducing, and other mental actions that contribute to the cognitive processing and understanding of information (Ford, 2004). Information behavior is adopted to embody both the seeking and use of information, stated as "the totality of human behavior in relation to sources and channels of information, including both active and passive information seeking, and information use" (Wilson, 2000, p. 49).

There is another concept often used interchangeably with ISB, i.e. information searching behavior. It is actually "the 'micro-level' of behavior employed by the searcher in interacting with information systems of all kinds" (Wilson, 2000, p. 49). Information searching has the model of information retrieval (IR) at its root. In coping with information overload, IR systems support information seeking based on the "query in, results out" mechanism (Rao, 2004). Web search engines are the most powerful IR systems, and they accommodate a considerable proportion of human information seeking activities. Besides searching, however, there are other approaches to discovering and finding information, e.g. browsing. Information searching behavior, for this reason, constitutes a sub-set of ISB.

The relationships between ISB and the above four related concepts can be represented with Figure 1. It expands Wilson's (1999) nested model to clarify the scope of ISB, helping define the

boundaries of this study.
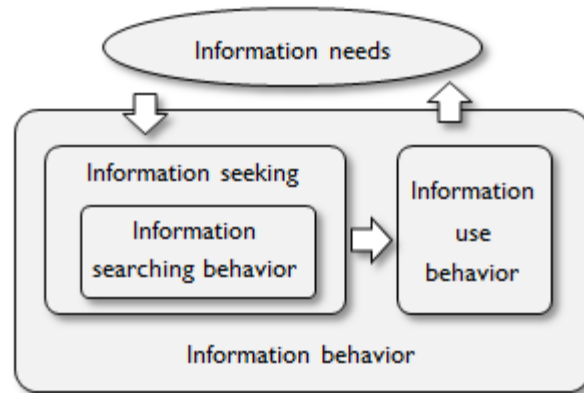
Figure 1. Relationships between ISB and related concepts

2.1.2 Theories and Models

ISB is an essential phenomenon in the LIS field. The understanding and explanation of this phenomenon are built upon various theories which are systems of assumptions, principles, and relationships. Most ISB theories are accompanied with explicatory models that play guiding and directing roles in the development of theories, especially "at the description and prediction stages of understanding a phenomenon" (Bates, 2005, p. 2-3).

Kim and Jeong's (2006) content analysis of 1,661 articles in 4 LIS journals from 1984 to 2003 indicated that the percentage of theory using papers increased continuously during this time period, whereas that of theory building papers declined recently, i.e. from 1999 to 2003. They detected the most evident decline in the areas of information seeking and use, and IR, suggesting that existing ISB theories have already been well established and extensively applied.

Prominent ISB theories fall into three general categories which are methodological,

empirical, and strategical. Dervin's (1992) sense-making theory and Wilson's (1999) problem-solving theory are the most famous ones taking the methodological perspective. Perceiving ISB as eliminating confusion or reducing uncertainties, they are conceptual tools of broad usefulness in understanding communication, information, and meaning (Tidline, 2005). Empirical theories come into being based on experiments and observations. Ellis (1993) studied academic researchers' behavior and developed a behavioral model addressing a series of characteristics underlying the complex information seeking patterns. Kuhlthau (1991) derived a model of information search process, incorporating the physical, affective, and cognitive aspects, from the investigation of common human experience. Both empirical models reside at the implementation level of information seeking.

Strategical theories are of the most interest to this dissertation study. Being more concrete and concentrated, these theories present information seeking strategies, the sets of ordered tactics that are consciously selected and applied to look for specific pieces of information (Marchionini, 1995). Judging by the degree to which an individual seeks information intentionally and positively, Bates (2002) differentiates four strategies – being aware, browsing, searching and monitoring – in her model of information seeking and searching (Figure 2a). They are also termed as passive attention, passive search, active search, and ongoing search respectively in the ISB component of Wilson's (1997) general model of information behavior (Figure 2b).

Searching, or active search, is the most familiar strategy to the majority of Web users. In contrast, browsing is just used to supplement searching in many cases, and being aware and monitoring are often not deemed as formal strategies. ISB researchers therefore center their

attention on the searching strategy while neglecting the other three in different degrees. Each of the four information seeking strategies, as a matter of fact, has been working for specific situations as the most appropriate one.
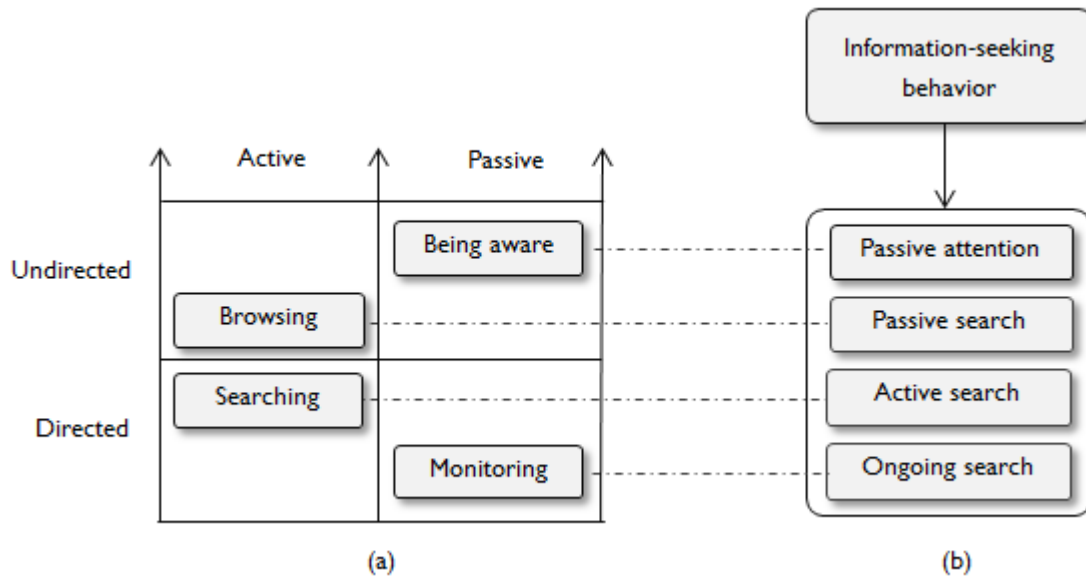
Figure 2. Models of strategical ISB theories: (a) Bates' model of information seeking and searching; and (b) ISB component in Wilson's general model of information behavior

Being aware, passive and undirected, is simply absorbing what randomly comes to us. There have been a handful of studies exclusively examining this information seeking strategy which is otherwise called "information encountering" (Erdelez, 1997; Williamson, 1998; Erdelez, 2004; Marshall & Jones, 2006). Although everybody encounters information, some users feel this happens very rarely to them while heavy encounters count on this strategy, and the rest are in between. Information can be encountered everywhere, such as in libraries, on the Web, or in the contact with other people. The Web may be the most conducive environment to information encountering because of the abundance of information available, but not for heavy encounters whose sensitivity

could easily lead to information overload. The information obtained this way can be used to address information needs arising in the past, at present, or in the future (Erdelez, 1999).

Still undirected, browsing usually starts with no particular goal either (Choo *et al.*, 1999). But browsers invest time and effort intentionally to acquire information or at least expect the acquisition of information, which makes browsing an active information seeking strategy. The reasons for browsing vary. It can help expand our knowledge to clarify our ill-defined information needs (Marchionini, 1995). Carefully organized information spaces encourage browsing, and unfamiliar or complex information spaces demand browsing so that we can gain an overview (Catledge & Pitkow, 1995). Browsing may be aroused by curiosity and often aim to discover and learn (Bates, 2002). Finally, browsers have lighter cognitive load than searchers since the former depend more on their perceptual abilities to recognize relevant information from the context (Marchionini, 1995).

In contrast, searching is basically driven by goals that can be clearly articulated. It is the most systematic information seeking strategy among the four, thus called the analytical strategy (Marchionini, 1995). At the beginning, searchers have to apply cognitive resources to recall from memory specific queries that express their information needs, but later they may be able to devote less attention and avoid disorientation or distraction because the search systems will automatically return the documents matching their queries. Nevertheless, searching can become very complicated, especially non-factual search in online systems like the Web. Searchers may have difficulties in selecting appropriate terms, differentiating synonyms and morphological variants, restricting the terms to specific search fields if there are any, so on and so forth. Searching hence

sometimes turns out to be a berrypicking process in which searchers' queries are continually refined with the information from previous querying attempts (Bates, 1989).

After we establish our basic framework of knowledge, ideas, beliefs or values with the above strategies, then monitoring, i.e. on-going search, will help us update or expand such a framework (Wilson, 1997). Monitoring is not opportunistic since we keep a back-of-the-mind alertness for topics that interest us and for answers to questions that we already have (Bates, 2002). Generally humans monitor because they do not feel an urgent need to spend an active effort. They would rather catch the useful information when it comes along. For some people such as academic researchers, however, monitoring is a necessity (Ellis *et al.*, 1993; Noble & Coughlin, 1997). They have to maintain current awareness of the developments in their fields through monitoring particular sources which can be electronic or printed materials or even their social networks (Bronstein, 2007).

2.2 INFORMATION SEEKING ON THE WEB

2.2.1 General Information Seeking

The availability of information on the Web during the 1990s has established Web documents as a popular medium through which users access information and search engines as a fundamental tool for information seeking (Tombros, *et al.*, 2005). As a consequence, the Web has become the primary

research context in which ISB is investigated (Martzoukou, 2005). Most existing ISB studies of Web users, in turn, concentrate exclusively on their interaction with a variety of search engines, i.e. information searching behavior (Yang, 2005).

In fact, the extraordinary research interest in searching is not that apparent in earlier studies which instead give more attention to Web browsing. Cockburn and Jones (1996) examined the usability of three Web client applications and identified their inadequacies in supporting user browsing, such as restricted page access and lack of page context. Tauscher and Greenberg (1997a; 1997b), collecting Web browsing data for over six weeks from 23 participants, reported that 58% of the pages visited by individual users are revisits. Byrne *et al.* (1999) made a task analysis with naturally-recorded verbal protocol data, providing a clear understanding of the tasks in which users engage during Web browsing and the time spent on those tasks.

These are in-depth ISB studies that have demonstrated empirical findings about Web users' browsing behavior. However encountering or monitoring behavior is scant of empirical works, only occasionally covered in comprehensive studies involving multiple information seeking strategies.

Catledge and Pitkow (1995) were the first to publish a major study of information seeking on the Web. Based on the analysis of a client-side log showing user navigation strategies and interface selections, the researchers distinguished three types of Web users: the serendipitous browsers (encountering), general purpose browsers (browsing), and searchers (searching). The lengths of their navigation sequences increase successively.

Choo *et al.* (1999) considered all four strategies in the investigation of 34 knowledge workers. Personal interview transcripts and WebTracker logs were used to identify 61 information

seeking episodes. 12 episodes were categorized as encountering, 18 as browsing, 23 as monitoring, and 8 as searching. In investigating 24 women in IT professions with the same methods, Choo and Marton (2003) identified 80 episodes. A total of 31 browsing episodes formed the largest group, followed by monitoring (21), encountering (14) and searching (14). These results contradict the prevalence of the searching strategy.

A more recent study by Huang *et al.* (2007), differently, characterized Web users' information behavior with three dimensions: "width" – number of categories of websites explored, "length" – number of site visited per category, and "depth" – number of pages downloaded per site. Their clickstream data analysis indicated that heavy Web users "are likely to explore more Web site categories, navigate more sites within a category, resort less to search engines in the navigation, consume more Web pages within a site, and consume them relatively quickly" (Huang *et al.*, 2007, p.1995).

While searching has not been found to be more important than other information seeking strategies in the above studies, we indeed see the dominance of research on information searching, probably due to the popularity of search engines among general Web users. By and large, the Web searching literature can be divided into two major streams: search log analyses that deal with explicit searching behavior, and user studies that reveal implicit factors contributing to the behavior (Martzoukou, 2005; Yang, 2005; Jansen & Spink, 2006).

2.2.2 Information Searching: Explicit Behavior

Search log analysis refers to the use of data collected in a search log to tackle particular research questions concerning the interactions among the searchers, the search engine, or the Web content during searching episodes (Jansen, 2008). And the search log is an electronic file kept on the server of a search engine and recording the interactions that have occurred during a searching episode between the search engine and its users (Jansen, 2008). Basic search log data include user IP address, query terms submitted by the user, and date and time of submission, whereas other types of data such as URL of the result page and URL of the page viewed may also be captured depending on the file format supported by the server (Jansen, 2006).

Search log analysis is a quantitative method applicable to various Web-based search environments. Jansen *et al.* (1998) and Silverstein *et al.* (1998) were among the earliest to conduct search log studies, with data from Excite and AltaVista respectively, two general Web search engines. Search systems for special purposes, including bibliographic tools (Blecic *et al.*, 1998; Wolfram & Xie, 2000; Bernstam *et al.* 2008), academic websites (Rozic-Hristovski *et al.* 2002; Wang *et al.*, 2003; Wolfram *et al.* 2009), digital libraries (Jones *et al.* 2000), and so on, have also been researched with this method. Language specific search systems, such as Chinese (Chau *et al.* 2007), Korean (Park *et al.* 2005), and Chilean (Baeza-Yates & Castillo, 2001) search engines, have attracted researchers' attention too. While many search log studies are based on relatively small sample sizes or short time lengths, or both, Jansen & Spink (2006) compared 9 search engines from the U.S. and Europe over a period of six years, presenting the most comprehensive breadth and depth of analysis in the

literature.

Guiding these studies is an established systematic search log analysis framework which consists of three levels: term, query, and session (Jansen, 2008). Most search log studies have made analysis at one or more of these levels.

2.2.2.1 Term

The term is the basic unit of analysis. Measures that can be examined at this level include term occurrence, total number of terms and unique terms, high usage terms, and term co-occurrence (Jansen, 2008). Of these measures, term co-occurrence is the most useful one. Ross and Wolfram (2000) analyzed the search subject content of Excite and categorized more than 1000 of the most frequently co-occurring term pairs into one of more of 30 developed subject areas. Their cluster analyses resulted in several well-defined high-level clusters of broad subject areas. In Huang *et al.* (2003), the researchers proposed a log-based term extraction and suggestion approach to interactive Web search. The approach could provide organized and highly relevant terms that co-occur in similar searches, and could exploit contextual information to make more effective suggestions.

2.2.2.2 Query

A query consists of one or more terms. Web users prefer short queries. The reported average query length never exceeded 2.5 terms, which was true to both English and non-English language search engines. (Jansen *et al.*, 1998; Beitzel *et al.*, 2004; Baeza-Yates & Castillo, 2001; Park *et al.*, 2005).

One-term queries are very common. Their percentages in the 9 search engines studied by Jansen and Spink (2006) ranged from 20% to 35%, and the percentage in Wang *et al.* (2003) reached 38%. If the first query fails to return satisfactory results, a user will submit subsequent queries which are usually different from the initial one (Jansen, 2006). It has been found by Rieh and Xu (2001) that while most query reformulation involves content changes, about 15% of the reformulation relate to format modifications. Another fact is that Web users also avoid complex queries that contain Boolean operators (Jansen & Spink, 2006; Beitzel *et al.*, 2004).

2.2.2.3 Session

It is not easy to define a session since the boundaries of a single search session will not be marked in search logs. One way of session detection is automatically grouping a user's consecutive queries on the same search topic into one session (He *et al.*, 2002). But its performance would be limited if users submit few queries and search on multiple topics (Özmutlu & Çavdur, 2005). The other way exploits the temporal characteristics of the queries. Two temporally adjacent queries submitted by a user belong to the same session only if their submission interval value is less than a cutoff value. The cutoff values vary from study to study, typically between 5 and 30 minutes (Huang *et al.*, 2001; Göker & He, 2002; Spink & Jansen, 2004; Baeza-Yates *et al.*, 2005).

Session length, i.e. the number of queries contained, and session duration, i.e. the total time the user spent interacting with the search system, are the two basic attributes of a session. Jansen *et al.* (2007) noted that on Dogpile.com the mean session length was fewer than 3 queries and the mean session duration was less than 30 minutes. The session-level analysis may also include the

click-through analysis which examines how users view the documents on result pages returned by the search engines. Based on the click-through data from AlltheWeb.com, Jansen and Spink (2003) found that more than 55% of all the users view only one document per query, and more than 66% of them view fewer than 5 documents in a given session. This echoes the previous finding that 85% of the time only the top 10 results were viewed (Silverstein *et al.*, 1999). Web users tend to evaluate their search results with the minimal effort, just like they do in constructing queries.

2.2.3 Information Searching: Implicit Factors

Although the above search log analyses offer an informative insight into Web users' searching behavior, they do not allow for an in-depth understanding of the behavioral patterns of individual users, because the user samples are anonymous and user involvement in those studies is zero. Such inadequacy has given rise to the other stream of information searching research which depends on user studies to explain what factors have induced the explicit behavior and how.

User studies of searchers may take place in natural environments that feature real information needs or in controlled environments where the search tasks are simulated. In either case, research data can be collected from the subjects in a many ways, such as questionnaires, interviews, observation, focus groups, think-aloud, and so on (Martzoukou, 2005). Their common purpose is to gather users' background information and elicit their thoughts and feelings attached to the search process. In previous studies, such data about the users has been analyzed mainly to reveal the impacts of four implicit variables which are search expertise, domain knowledge, cognitive

styles, and affective characteristics.

2.2.3.1 Search Expertise

The recognition of search expertise as a major factor determining search performance may be traced back to early 1990s, even to the 1980s. Experienced searchers have been found to locate information more quickly and make fewer errors than novices in traditional IR systems (Fenichel, 1981; Marchionini *et al.*, 1990). Although users' Web expertise is also significantly related to search success (Yee *et al.*, 1998), their search expertise is still a better predictor (Kim, 2001). Sutcliffe *et al.* (2000) divided a group of medical student searchers into good and poor searchers according to their search outcomes. The researchers came to the conclusion that poor searchers used simpler queries with fewer terms, and gave up more easily, while good ones iterated more frequently and evaluated search results more carefully.

In addition, Navarro-Prieto *et al.* (1999) was interested in the interaction between search experience and task type. In the fact-finding task, experienced participants were more confident in choosing search keywords, but novice participants instead were more influenced by external representations in the search result pages, e.g. using the words from certain result items to improve their queries. In the exploring task which demanded for a nonlinear search process, experienced participants again proceeded in a more structured way with novice participants having no planning in advance.

2.2.3.2 Domain Knowledge

Experts in specific domains will be able to use a variety of terminologies in their queries and spend less time in viewing the search results. The level of searchers' domain expertise therefore is no less influential than their search expertise in query formulation and modification as well as in search result evaluation. At the same time, these two types of expertise have shown independent and combined effects. In Hölscher and Strube (2000), participants possessing high expertise in both search and domain knowledge were overall the most successful in searching. Even if deficits presented in one type of expertise, they could be compensated by the other. "Double novices", unfortunately, featured the highest proportion of query reformulations and the smallest number of result documents for closer examination. What's worse, the changes they made to their queries were usually ineffective, and most documents they viewed turned out to be irrelevant. Jenkins *et al.* (2003) obtained similar findings when investigating four groups of nurses with different combinations of expertise. Specifically, domain expertise determined result evaluation criteria, and search expertise affected the scope of search space.

2.2.3.3 Cognitive Style

Cognitive style refers to the individual's characteristic way of organizing and processing information (Goldstein & Blackman, 1978). There are two major types of cognitive styles: field dependence (FD) and field independence (FI). While FD individuals are more likely to be dominated or influenced by the surrounding perceptual field, FI individuals are adept at overcoming such influences (Kim & Allen, 2002). In Web searching, the latter tend to find information more efficiently than the former,

with shorter search duration and fewer search steps (Palmquist & Kim, 2000; Wang *et al.*, 2000). Cognitive style itself actually does not directly act on the search process, instead, the level of users' online search experience determines the way their cognitive styles influence their search performance. Kim (2000) explored the interaction between cognitive style and search experience and ascertained that the differences brought about by participants' cognitive styles disappeared in those who had adequate online search experience, implying that FD searchers would conquer their inefficiencies in searching by gaining search experience.

2.2.3.4 Affective Characteristics

Users may have different emotions, such as relaxation, anxiety, and frustration, at different stages of a search process. Nahl (2005) measured several affective variables in the form of rating scale filled out by college students at the beginning and end of weekly Web search sessions throughout a semester. The results showed that "affective coping skills" consisting of self-efficacy and optimism had a positive impact on search performance and that higher affective skills could compensate for low cognitive skills. In a recent study on affective characteristics, Kim (2008) assigned one general task and one specific task to 67 undergraduates. The effects of emotion control were found significant on users' searching behavior, but not on their performance. Participants with low emotion control are liable to make frequent and quick search moves, and especially in the general task which is more complex, they experience a high level of uncertainty.

2.3 SOCIAL TAGGING SYSTEMS

2.3.1 Background: Web 2.0

It has been a decade since information architect Darcy DiNucci (1999) coined the term "Web 2.0" for her nascent vision of the future Web. Despite of the vast amount of citations today, there still lacks a widely agreed definition for this term. What the literature frequently refers to is a set of core principles of Web 2.0 put forward by Tim O'Reilly (2005). An important part of his description about Web 2.0 is the positioning of the Web as a platform for users to participate and collaborate.

In the Web 1.0 era, most websites are merely information sources which rely on site owners to provide information and site designers to organize and present information, with users only able to retrieve, view, and download information through a Web browser. Web 2.0 sites, in contrast, offer tools and services in addition to information (Solomon & Schrum, 2007). Users are enabled to run a variety of Web-based software applications entirely in a browser, e.g. word processor and photo editor. Without installing any software tools on a local disk, they do not need to work on a specific computer; and Web services are responsible for keeping their files and projects that thus can be accessed from any computer. In this way, the Web serves as the workspace where users complete many regular tasks with great mobility (Bradley, 2007).

As further interpreted in O'Reilly (2005), this is a participatory platform since Web 2.0 encourages user participation. On the one hand, users are allowed to own data on Web 2.0 sites. They have free control over their data, though there can be privacy and copyright issues. On the

other hand, they are also allowed to employ lightweight technologies, namely, XHTML, XML, CSS, JavaScript, AJAX, etc., to customize the applications to suit their needs. Such client-side programmability both means richer experiences to the users and added value to the applications.

Collaboration is an even more central concept of Web 2.0. It's believed that the survival of such leading sites as Yahoo!, Google, eBay, and Amazon in the transition from Web 1.0 to Web 2.0 should be attributed to their efforts in "harnessing collective intelligence" (O'Reilly, 2005). As revisited several years later, the notion of "harnessing collective intelligence" is a fundamental idea underlying successful Web 2.0 sites (O'Reilly & Battelle, 2009). Believing that "a large group of people can create a collective work whose value far exceeds that provided by any of the individual participants", they are constructed in such a way as to direct their users to perform specific tasks, including building an online encyclopedia, adding data points onto a map, and finding the most popular news stories, etc. (O'Reilly & Battelle, 2009, p. 2).

The above-mentioned websites, old or new, and hundreds of thousands of other sites with the "Web 2.0" mark, are built upon powerful social software to assist users to collaborate in one way or another. Social software comprises a range of easy-to-use Web-based software applications. Instant messaging, electronic mailing lists, and Internet forums are the traditional forms which have been around for decades. Modern ones mostly emerge within the past 10 years, including blogs, wikis, social network services, social bookmarking, podcasting, and virtual worlds, just to name a few (Farkas, 2007).

Social software applications are also called "social media", an alternate however emphasizing their communication purposes as media and the online content they transmit (Newson *et al.*, 2008).

In fact, what turn the Web into a collaborative platform are the abilities of social media to help individual users establish and maintain relationships with others by overcoming spatial and temporal barriers, and then to help linked users leverage existing relationships to find and create knowledge (Burkhardt, 2009). By considering both aspects, social media efficiently reproduces the offline scenarios found in social science studies: people heavily rely on their social ties in everyday learning, information seeking, decision making, problem solving, so on and so forth (Cross & Parker, 2004).

2.3.2 Social Tagging System: The Three-Part Architecture

With different functionalities, social software applications facilitate user collaboration in different manners. For example, instant messaging, social network services, and virtual worlds specifically concentrate on connecting people and building online social ties (Kroski, 2005). The relationships between users are usually mutually acknowledged friendships which involve them in chatting, networking, gaming, and other straightforward social activities (Jiang & He, 2007). In certain sense, they are people-oriented tools for explicit interaction.

User interaction, indeed, can also take the "implicit" form, as seen in social bookmarking, social libraries, multimedia sharing and other information-oriented tools (Mieszkowski, 2005). That is to say, making friends, though still supported, is secondary to information related activities — storing, tagging, sharing, and discovering information resources, being them bookmarks, or photos, etc. In spite of no direct interaction, users collaborate on the bibliographic records for their collected resources in the form of automatic aggregation of individual actions so that the best or

most relevant information comes to the surface (Porter, 2008).

This latter stream of social software applications is often referred to as social tagging systems in general. Being information-oriented, social tagging systems are of particular interest to this dissertation study because of their significance to LIS. For one thing, they have undoubtedly produced many information repositories of great value by gathering copious digital resources that may not be found elsewhere on the Web. For another, by depending on users to assign meaning to resources in the form of tags, social tagging systems have brought about a momentous revolution in modern cataloging and classification.

The best way to understand social tagging systems is through their universal architecture which embodies three elements – resources, users, and tags, as in Figure 3 (Smith, 2008). There exists no one-to-one mapping between any pair of these elements (Marlow *et al.*, 2006). A particular resource may be tagged by many uses and have many tags; a particular user may create many tags and tag many resources; a particular tag may be used by many users and attached to many resources. Below a closer look will be taken at each of these elements.
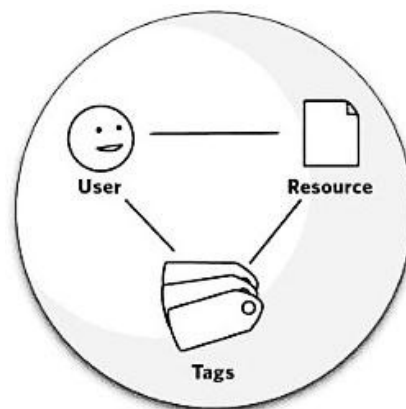
Figure 3. Social tagging system architecture (Smith, 2008, p.4)

2.3.2.1 Resources

Practically everything, physical or digital, can be tagged. What exactly is being tagged in social tagging systems? As already mentioned, there are digital resources including the actual photos, videos, and audios, etc. in multimedia sharing systems. Other categories of social tagging systems, in contrast, can just accommodate the pointers (e.g. bookmarks) to the original objects (e.g. webpages) (Smith, 2008). Where are these resources derived from? Original digital resources (e.g. family videos) are mainly contributed by users who are also probably the creators, while pointers stand for open resources (e.g. music videos) that exist either globally on the Web or locally in the systems (Kalbach, 2007).

The type and source of a resource will affect who tags it and how it is tagged (Marlow *et al.*, 2006). The tagging of bookmarks, citations, and other pointers is collaborative since a lot of people can describe the same resource with their own tags. The system is able to recognize the resource by its public identification and aggregate all the tags applied to engender one bibliographic record for it. When it comes to user-contributed resources, however, tagging right is usually limited to those specific users who upload them. The system may still aggregate tags across users, but regardless of resources. This is not collaborative tagging, strictly speaking (Golder & Huberman, 2006).

2.3.2.2 Users

Users are the most vibrant element among the three, because they decide what resources to tag and what tags to use. User tagging can be motivated by several incentives. Tagging offers individual users self-organization and retrieval that are immediate personal benefits (Terdiman, 2005). Tagging in

shared space may further generate network effects, offering classification and discovery for known or unknown audience (Trant, 2009; Kroski, 2005). Also, tagging can also work as a self-presentation tool for users to show activism and attract attention, and as a self-communication channel for them to express opinions liberally (Zollers, 2007; Marlow *et al.*, 2006).

Users' knowledge, experience, and judgment have a great impact on the way they tag. Inconsistencies in these aspects thus lead to deficiencies in the result of social tagging, of which the vocabulary problem and basic level problem are the two major ones (Golder & Huberman, 2006). The former has more to do with the users' understanding of the semantic relations between words and their referents (Furnas *et al.*, 1987). In tagging the same resource, users may make different choice of words due to polysemy, synonymy, homonymy, and plural. Reflecting users' cognitive habits of hierarchy and categorization, the latter problem is that related terms vary along a continuum of specificity (Tanaka & Taylor, 1991). Whether a user will tag a resource with a superordinate term or a subordinate one is directly relevant to his or her interaction with it.

In addition to tagging, most social tagging systems also value user connectivity that is a powerful feature for encouraging participation (Marlow *et al.*, 2006). There are two types of connections between users: one-way and reciprocal (Smith, 2008). Sometimes both of them are allowed in one system. These connections constitute system-wide social networks which are the foundation of social navigation – navigation that is driven by the actions of other people (Svensson *et al.*, 2001). Groups, by gathering similar users around particular topics, is an additional means of increasing social connectivity, though involving no explicit connections.

2.3.2.3 Tags

Social tagging is in essence a process of subject analysis (Schwartz, 2008). As a part of cataloging, subject analysis deals with "determining what the intellectual content of an item is 'about', translating that 'aboutness' into the conceptual framework of the classification or subject heading system being used, and then translating the conceptual framework into specific classificatory symbols or specific terminology used in the classification or subject heading system" (Taylor, 2004, p. 275). Such conventional practice is only partially applicable to social tagging, however. Although tags describe aboutness, they are not translated from the conceptual framework of any preexisting classification or subject heading system. Instead, they are aggregated to engender social classifications, i.e. folksonomies, in a bottom-up fashion (Quintarelli, 2005).

Folksonomies, exactly speaking, are name spaces without rigid hierarchy or exclusive categories (Hammond *et al.*, 2005). For lack of integral structures, they are very responsive to changes and able to keep updated all the time, which is impossible in established taxonomies. Another widely agreed strength of folksonomies is that they are much less expensive to create than taxonomies built by professionals as the money and time costs have been distributed among a large number of users (Chi & Mytkowicz, 2006). The other side of the coin is that the findability of folksonomies is low because different people use different tags to describe the same thing (Morvill, 2006).

For users to make better use of folksonomies, most social tagging systems present them with tag cloud visualizations where tags are often displayed in alphabetical order and the font size

implies tag use frequency (Sinclair & Cardew-Hall, 2008). Sometimes other attributes such as text weight and color may also be adopted to highlight certain features. But on the whole tag clouds are based on simple visualization techniques and very easy to use. A click on the tag of interest in a cloud will redirect users to all the resources associated with it and their taggers, and the related tag(s) as well. The same end can be achieved by searching with that tag since most system search engines support queries that are single tags or tag combinations. But browsing the folksonomy is more like exploration than known-item searching, as noted by Winget (2006) in a study of the ways that tags support information discovery within Flickr.

3.0 METHODOLOGY

This dissertation study employed both quantitative and qualitative methods. Section 3.1 describes the overall mixed methods research design, including the reasons for choosing such a design, the way it applies to this study, and the challenges it brings about. Then the research setting, a representative social tagging system, is introduced in Section 3.2, with a focus on its major categories of webpages. Section 3.3 elaborates on the data collection and analysis processes of this study. It breaks down into two parts respectively detailing the quantitative phase consisting of a clickstream data analysis and an online survey and the qualitative phase involving a focus group. The largest portion of the section is devoted to the clickstream data analysis for it being the central method of this study and never implemented in previous research. Finally, Section 3.4 discusses the methodological limitations that may affect the research validity.

3.1 RESEARCH DESIGN

There are three types of research designs: quantitative, qualitative, and mixed methods (Creswell, 2009). While quantitative research tests objective hypotheses by measuring the relationships among

variables in numerical ways, qualitative research explores and interprets socially constructed realities (Newman *et al.*, 2003). They represent the different ends of an "interactive continuum", in the middle of which is the mixed methods research (Newman & Benz, 1998). This dissertation study, in particular, adopts the mixed methods design that "combines the qualitative and quantitative approaches into the research methodology of a single study or multiphased study" (Tashakkori & Teddlie, 1998, pp. 17-18).

Any method has its inherent limitations, but seeking convergence across quantitative and qualitative methods provides the opportunity to compensate for the weaknesses of each method (Creswell, 2009). The mixed methods approach, on the one hand, allows the researchers to understand the phenomenon from multiple angles, which help them extract more information from the underlying data and generate more meaning. On the other hand, it offers the researchers more than one way to assess their findings, thus reducing possible biases and ensuring the validity of data interpretation. These are the two major rationales for mixed methods: representation and legitimation (Onwuegbuzie & Teddlie, 2003).

The mixed methods research design takes four different forms: triangulation, embedded, explanatory, and exploratory (Creswell & Plano Clark, 2007). The first two are also known as concurrent or parallel designs in which quantitative and qualitative data are collected at the same time and integrated in the interpretation of the overall results (Creswell *et al.*, 2003). The other two types, differently, are based on the sequential mixed model in which "multiple approaches to data collection, analysis, and inference are employed in a sequence of phases" (Tashakkori & Teddlie, 1998, pp. 149-150).

Being sequential, this study was composed of two phases. The first phase mainly included a clickstream data analysis addressing the quantitative research questions, RQ 1 and RQ 3. It was complemented by an online survey conducted especially for RQ 2, the follow-up question of RQ 1. This phase reflects one of Morse's (2003) deductive research programs, i.e. composed of two quantitative methods, one of which is dominant. During the second phase, a focus group discussed in depth the major findings from the prior phase. Such research design, exactly speaking, is explanatory sequential. The researcher started with the collection and analysis of quantitative data, and the subsequent collection and analysis of qualitative data enabled the researcher to interpret the quantitative results more accurately. This should be distinguished from the exploratory sequential design in which the qualitative method precedes and helps develop or inform the quantitative method whose variables, measures or instruments may not be determined yet (Creswell 2009).

More precisely, this study was grounded upon the follow-up explanations model of the explanatory sequential design, which attached more importance to the quantitative phase. "The researcher identifies particular quantitative findings that need additional explanation, such as statistical differences among groups, individual who scored at extreme levels, or unexpected results. The researcher then collects qualitative data from participants who can best help explain these findings" (Creswell & Plano Clark, 2007, pp.72). With the notations that originally appeared in Morse (1991), the overall research design of this study can be represented as "(QUAN + quan) → qual". The plus sign (+) means concurrence, the arrow (→) means sequence, and uppercase indicates dominance.

Although the follow-up explanatory sequential design is the most straightforward one

among all the mixed methods designs, this study was still confronted with several major challenges. The first and foremost challenge was the high consumption of time for completing two phases. Out of this consideration, the researcher not only budgeted sufficient time for the entire study, but also used more manageable sample sizes for both phases. The next challenge came forth in the stage of sample selection. Different methods have different sampling requirements and procedures. In order to keep a certain degree of consistency in sampling, the clickstream data analysis, online survey and focus group drew appropriate individuals from the same population separately. Finally, there existed an intrinsic challenge due to the sequence of two research phases. The researcher had to wait until the quantitative phase generated initial findings, and then made a clear plan of how focus group participants would be selected and investigated for the qualitative phase.

3.2 RESEARCH SETTING

Douban (http://www.douban.com/) is one of the largest social tagging systems on the Web. A Chinese-language site founded in 2005, it has attracted more than 46,000,000 registered users from all over the world. To be more specific, Douban is a social library system for users to discover and collect three types of resources – books, movies, and music albums, store them all in one personal library, and share their libraries with others. Similar English-language systems include LibraryThing, IMDb, and Discogs as mentioned in Section 1.1. As a typical social tagging system, Douban encourages users to assign tags to the resources at the time of collecting and allows them to edit or

delete the tags later. What's special about Douban is that the type of a resource determines the type of its associated tags. Hence, there are book tags, movie tags, and music tags in the system, with each type of tags being an independent folksonomy. This is a main difference between Douban and most other social tagging systems, each of which has one single folksonomy.

The top half of Figure 4 illustrates the website architecture of Douban from the perspective of information organization. To differentiate the three types of resources, Douban divides into the Book[19], Movie[20], and Music[21] sub-sites. They appear in the global navigation bar together with the website home. If visited by anonymous users, the Douban home page will simply exhibit the thumbnails of the top resources of each type, and the three second-level home pages will present recent resources/reviews, popular resources/reviews, frequently used tags, highly recommended "doulists" – user-compiled lists of similar resources, and "Douban250" – 250 resources with the greatest popularity and highest ratings. For registered users who sign in, the home pages will turn into "mine" pages on which additional recommendations of resources generated by the system based on their current collections or made by their contacts will be available for viewing. As a whole, Douban users can conveniently start with exploring what most people are interested in or what the system has matched for their tastes, if they do not bother to look for resources.

Douban has a flat and wide hierarchy, just like other social tagging systems. Immediately below the sub-sites are the folksonomies, i.e. the whole range of resource categorizations in the form of tags. A tag page aggregates the tag's associated resources, arranges them in a descending

---

[19] http://www.douban.com/book/, currently http://book.douban.com/
[20] http://www.douban.com/movie/, currently http://movie.douban.com/
[21] http://www.douban.com/music/, currently http://music.douban.com/

order in terms of tag usage frequency, and provides navigation to relevant tags which are frequently

co-assigned with that tag to these resources. Millions of resources lie at the bottom level, with their

non-exclusive belongingness to the above categorizations permitted. If a resource is never tagged, it

belongs to none of the categorizations. Each resource page contains four basic parts: (1) original

information about the resource, such as creator, date, and brief introduction; (2) system-gathered

history data, e.g. the most frequently attached tags, often co-collected resources, and resource

collectors, etc.; (3) user-contributed content, including reviews, doulists, and discussion topics; and

(4) rating and collecting the resource – users will complete the action in a pop-up window with the

option of adding tags.
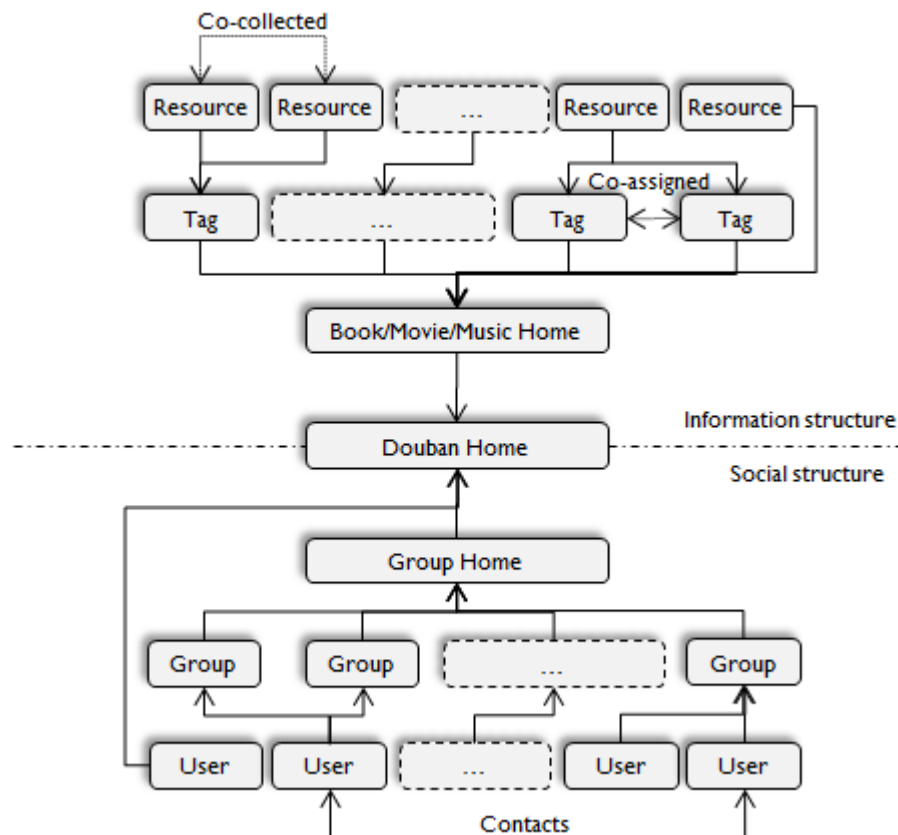
Figure 4. Douban's website architecture consisting of an information structure and a social structure

Emphasizing the secondary role of Douban as an online community, its social structure can be read from the bottom half of Figure 4. At the lowest level we see the users. They may share similar interests and constitute various groups, such as "Grammy"[22], "trip"[23], and "Macintosh"[24], just to name a few. Every group has a discussion forum welcoming any topic pertinent to its theme, as well as a collection of related resources added by authorized group members. Now Douban is the home of 170,000 interest groups classified under its Group[25] sub-site. This social architecture, however, is not strictly hierarchical. On the one hand, a considerable proportion of all the users never join any groups, so they are directly subordinate to the top level, i.e. the Douban home. On the other hand, users add one another as contacts explicitly, resulting in a network structure.

Interest groups and social networking are common features seen in many social tagging systems, such as CiteULike, Discogs, and Librarything. Despite their enthusiasm for affiliating to groups and building relationships, users usually still center their activities around resources. One may interact with a particular resource in several ways, such as collecting it, writing a review about it, adding it to a doulist, recommending it to other users, and so on. Each user page therefore mainly functions as a showcase of the individual's collected resources, contributed reviews, doulists, and recommendations, etc. Next, it serves the social networking purpose by showing the individual's contacts and groups if there are any as well as diaries, photos, blog posts, and messages which he or she wants to share.

---

[22] http://www.douban.com/group/Grammy/

[23] http://www.douban.com/group/trip/

[24] http://www.douban.com/group/macintosh/

[25] http://www.douban.com/group/

To sum Figure 4 up, Douban accommodates five essential clusters of webpages: home pages ("mine" pages when users sign in), resource pages, tag pages, group pages, and user pages. Thanks to plenty of hyperlinks, users are able to navigate smoothly within each cluster or across the clusters to browse linked pages, except for from the tag cluster to the user cluster. Hyperlinks sometimes lead to disorientation however, for which reason Douban provides an internal search engine. Users can perform general searches that return all the matching resources or narrow the searches down to one specific resource type, but their queries are restricted to resource title, creator, and coding (i.e. ISBN, IMDB, or ISRC).

Nevertheless, there is also a noticeable cluster of peripheral pages beyond the fundamental website architecture of Douban. This social library system has developed into a multifunction social media where people are also supported to launch online and offline events, publish weblogs, share photos, communicate via Douban's email service, and so on. These activities will not lead users to new resources, but they are interwoven with the information seeking activity. For instance, a resource page shows the events that the resource collectors have participated in, or a user page displays the photo albums of that user, and it is normal that one wants to view the event or photo pages to satisfy his or her curiosity. Usually, removing peripheral pages from a visit has no influence on the results of information seeking.

In this dissertation study, the goal of information seeking in the context of Douban is defined as finding a book, a movie, or a music album of interest, i.e. finding an interesting resource regardless of type. Every time a user reaches a resource page from any other page, we will say that the user finds a resource. After viewing the details about the resource on that page, the user may

decide whether to add it to library, which signals whether he or she is interested in it. Resource finding and collecting are the focus of research. It should be mentioned that resource collecting in Douban is made a little complicated since users have to select one of the three tenses – present ("is reading / watching / listening to"), future ("wish to read / watch / listen to"), and past perfect ("have read / watched / listened to") – in order to indicate their familiarity with the resources. Such factor is not considered in this study, given that a resource, once collected, must be interesting to the collecting user no matter which tense is selected.

3.3 DATA COLLECTION AND ANALYSIS

3.3.1 Clickstream Data Analysis

Clickstream data analysis, the primary method in the quantitative phase of this dissertation study, was conducted with a transaction log file from Douban's server. Simply speaking, a clickstream is a click path. It provides information about the sequence of webpage requests made by a user when navigating through a website (Montgomery, 2001). Between the user entering and exiting the website, every time he or she clicks a link or a button on a webpage, the clicking will be recorded. One approach to recording such data is client-based, known as page tagging, which runs a JavaScript code embedded in each webpage to record successful page requests and sends the information back to a remote server. There is also a server-based approach – a transaction log is created on the

website server to record all the requests for pages and other resources, both successful and unsuccessful (Jansen, 2008). This study chose the latter approach which has been deemed the least expensive and most unobtrusive means to collect clickstream data from a large number of users (Jensen & Pooch, 2000).

Clickstream data analysis has been a principle form of Web analytics, the measurement of user behavior on a website (Booth and Jansen, 2008). It is however seldom employed to study Web users' information seeking behavior, instead mostly seen in e-commerce research. Online retail stores have been making extensive analysis of their customers' clickstream data to reveal website navigation patterns, and to evaluate the effectiveness of marketing and merchandising efforts (Nasraoui *et al.*, 2003; Moe & Fader, 2003; Lee, 2001; Chatterjee *et al.*, 2003). By examining what products the customers viewed, how many clickthroughs the advertisements generated, whether the customers completed the purchases, and so forth, e-commerce website designers are able to identify profitable and unprofitable site features.

In a similar manner, the clickstream data of an information-rich website informs what information resources the users are interested in, how they look for the resources, and whether they find their needed resources. These are important clues which can be utilized to improve the overall information seeking experience within the website. Introducing the clickstream data analysis to ISB research, in a certain sense, will dwarf the prevalent search log analysis as reviewed above. Search logs are unique to search tools and limited to capturing users' searching behavior, such as submitting queries and viewing search results, while clickstreams are universal log data in all kinds of information seeking environments, including social tagging systems which demand for multiple

information seeking strategies besides searching to cope with the complexity of Web 2.0 realities.

3.3.1.1 Data Collection

A random transaction log file was directly requested from Douban. It contains around 20 million user clickstream records generated on the Web server over a 24-hour period, from 00:00:00 to 23:59:59 on December 12, 2008. Websites are usually very careful about releasing their transaction logs for fear of offending users' privacy. Douban also gave full consideration to this issue and had a technician encrypt all the users' identities in the log file. Specifically, each user was assigned a new ID, i.e. a string of digits that assumed no meaning but helped the researcher identify a specific user. According to Douban, the log file originally followed the W3C Extended Log Format, one of the common formats today (Booth & Jansen, 2008), but for quicker processing it was simplified to five basic fields – *USER ID*, *REQUSTED_URL*, *METHOD*, *REFERRING_URL*, and *TIME*.

USER ID   User's IP address (for users who did not sign in) or user name (for users who signed in) disguised with a 9 or 10-digit number which can be positive or negative;

REQUSTED_URL   URL of the page requested by the user. The actual page can be accessed by typing "http://www.douban.com" + "URL" in a Web browser, also applicable to the *REFERRING_URL* field;

METHOD   Type of the request. The majority of the records in this log file are associated with the "GET" method, i.e. requesting a page from the Web

server. Also seen is the "POST" method that modifies the content of the data

stored on the server;

REFERRING_URL   URL of the page from which the user access the page in the corresponding

REQUSTED_URL field;

TIME   Exact time when the user makes the request and displayed in the AM/PM

format.

USER ID	REQUSTED_URL	METHOD	REFERRING_URL	TIME
2061537704	/people/rink/	GET	-	8:43:51 PM
2061537704	/j/subject_suggest?q=%E9%87%91%E5%AD%97%E5%A1%94	GET	/people/rink/	8:44:00 PM
2061537704	/j/subject_suggest?q=%E9%87%91%E5%AD%97%E5%A1%94%E5%8E%9F%E7%9	GET	/people/rink/	8:44:05 PM
2061537704	/subject_search?search_text=%E9%87%91%E5%AD%97%E5%A1%94%E5%8E%9	GET	/people/rink/	8:44:05 PM
2061537704	/j/subject_suggest?q=%E9%87%91%E5%AD%97%E5%A1%94%E5%8E%9F%E7%9	GET	/people/rink/	8:44:08 PM
2061537704	/subject/3189420/?i=0	GET	/subject_search?search_text=%E9%87%91%E5%AD%97%E5%A1%94%E5%8l	8:44:10 PM
2061537704	/subject/3189420/?i=0	GET	/subject_search?search_text=%E9%87%91%E5%AD%97%E5%A1%94%E5%8l	8:44:35 PM
2061537704	/j/subject/3189420/interest?interest=collect&rating=5	GET	/subject/3189420/?i=0	8:44:43 PM
2061537704	/j/subject/3189420/interest	POST	/subject/3189420/?i=0	8:45:01 PM
2061537704	/subject/3189420/	GET	-	8:45:01 PM
2061537705	/group/topic/1865987/	GET	http://www.baidu.com/s?wd=%B6%F5%C2%D7%B4%BA%D0%A1%B3%AA	8:41:13 PM
2061537705	/group/topic/4249302/?start=100	GET	/group/topic/4249302/?from=mb-86987056	8:41:42 PM
2061537705	/group/topic/4249302/?start=200	GET	/group/topic/4249302/?start=100	8:50:33 PM
2061537706	/	GET	-	8:43:26 PM
2061537706	/subject/1427083/?rec=V&rec=V	GET	/	8:48:05 PM
2061537706	/doulist/188962/	GET	/subject/1427083/?rec=V&rec=V	8:48:23 PM
2061537706	/subject/1721591/	GET	/doulist/188962/	8:48:30 PM
2061537706	/book/	GET	/subject/1721591/	8:48:44 PM
2061537706	/book/tag/%E5%93%B2%E5%AD%A6	GET	/book/	8:48:49 PM
2061537706	/book/tag/%E5%93%B2%E5%AD%A6?start=20	GET	/book/tag/%E5%93%B2%E5%AD%A6	8:49:27 PM
2061537706	/book/tag/%E5%93%B2%E5%AD%A6?start=40	GET	/book/tag/%E5%93%B2%E5%AD%A6?start=20	8:49:55 PM
2061537706	/book/tag/%E5%93%B2%E5%AD%A6?start=60	GET	/book/tag/%E5%93%B2%E5%AD%A6?start=40	8:50:17 PM
2061537706	/book/tag/%E5%93%B2%E5%AD%A6?start=80	GET	/book/tag/%E5%93%B2%E5%AD%A6?start=60	8:50:32 PM
2061537706	/book/tag/%E5%93%B2%E5%AD%A6?start=100	GET	/book/tag/%E5%93%B2%E5%AD%A6?start=80	8:50:51 PM

Figure 5. A snippet from the original transaction log file provided by Douban

The CSV-formatted log file received from Douban was imported into a single table named

original_data in Microsoft Access. Figure 5 captures a snippet from this table after being sorted by

USER ID firstly and *TIME* secondly, containing 24 clickstream records belonging to 3 users. It should

be mentioned that Chinese tags or search keywords in the *REQUSTED_URL* and *REFERRING_URL*

fields are not directly readable in the log file which is based on the UTF-8 encoding scheme, as can

be seen in the above table. Since this study involves no semantic analysis, they were not translated into Chinese characters during the data processing.

3.3.1.2 Data Cleaning

Data cleaning is important to any types of Web analytics – "the discovered associations or reported statistics are only useful if the data represented in the server log gives an accurate picture of the user accesses to the Web site"(Cooly, *et al.*, 1999, p. 12). In this clickstream analysis, the first cleaning step was removing corrupted records from Table *original_data*. They were the erroneous data produced when the Web server logged the data incorrectly and could be easily recognized by sorting each field in sequence. Errors usually appear on the top of, bottom of, or grouped together in the sorted column because they do not fit the pattern of the normal data in the same column (Jansen, 2006).

In addition to corrupted records, there was a considerable amount of redundant data in Table *original_data*. They failed to reflect users' information seeking behavior within Douban. Filtering them out would minimize the size of the table and expedite the analysis. The major types of irrelevant records eliminated included:

(1) External links (both inbound and outbound): *REQUSTED_URL* or *REFERRING_URL* begins with "http://" or "/ninetaps" (Douban's affiliated blogging service);

(2) Requests with unknown referring pages: *REFERRING_URL* = "-";

(3) Requests that refresh the current pages: *REQUSTED_URL* = *REFERRING_URL*;

(4) Requests by Web search engine robots or crawlers (mainly Googlebot): *REFERRING_URL* = "www.google.com";

(5) Requests resulting in the modification of content: *METHOD* = "POST";

(6) Requests triggered by a JavaScript action except for resource collecting: *REQUSTED_URL* begins with "/j/" but not with "/j/subject/";

(7) Requests for pictures, styles, scripts, and other resources: *REQUSTED_URL* or *REFERRING_URL* ends with ".jpg", ".gif", ".png", ".css", ".asp", and ".js";

(8) Requests for Douban services[26] – Widgets (exhibiting users' Douban libraries on their external blogs), RSS feeds (subscribing to Douban content), Bookmarklet (bookmarking external webpages on Douban), ISBN search (searching for books with their ISBN identifiers), and API (interacting with Douban data or functionalities in external websites or programs): *REQUSTED_URL* or *REFERRING_URL* begins with "service/badge", "/feed", "/service/bookmarklet", "/isbn", "/service/api" or "/service/auth".

After data cleaning, 10,303,684 clickstream records remained in the table which was renamed *cleaned_data*. The entire *METHOD* column was deleted for displaying one invariable value – "GET". The 4 fields left were *UID* (originally *USER ID*), *REQ* (originally *REQUSTED_URL*), *REF* (originally *REFERRING_URL*), and *TIME*. Table *cleaned_data* contains 269,658 distinct users. 22% (N = 59,356) of them have only 1 record each, 69% (N = 186,914) 2 to 99 records, and 9% (N = 23,388) 100 records or over. At the higher end, there are 638 extreme users, each of whom has no

---

[26] http://www.douban.com/service/

less than 1,000 records, and the maximum number of clickstream records a user may have is 27,050.

3.3.1.3 Data Analysis

Analyzing clickstream data was unprecedented in ISB research, so there was no readily usable method. The popular search log analysis framework, i.e. studying search logs at term, query, and session levels, was obviously not applicable here. However, a promising framework especially for clickstream data analysis has been proposed by Sen *et al.* (2006) though never implemented, introducing three behavior tracing concepts – *footprint*, *track*, and *trail*. A footprint represents a single clickstream record created by the interaction between the user and a webpage, and a collection of footprints constitute a track. If the tracks of a group of users are similar, they cluster into a trail, comprising similar behaviors, attitudes, beliefs, and values.

Taking into account the characteristics of real-world clickstreams, the researcher established a better suited analysis framework composed of three levels – footprint, movement, and track. Most dictionaries define the term "footprint" as an impression left by a foot while "movement" an act of changing location or position. A footprint, so to speak, is the result of a movement. The latter dynamic concept, *movement*, is apparently more appropriate for representing a single clickstream record that describes a certain user changing his or her location from a referring page (in the *REF* column) to a requested page (in the *REQ* column) at a certain time point. Instead, *footprint* in this new context particularly refers to the requested page of a clickstream record, the resulting status of the interaction between the user and the referring page which in turn is the footprint of the last

record. Correspondingly, the meaning of *track* needs to be modified to a series of consecutive movements. It provides a user's navigational history when visiting a website.

The concept of trail introduced by Sen *et al.* (2006) was excluded from this clickstream data analysis framework due to its limited applicability to e-commerce websites. In a travel website, for example, we see air-trail for purchasing flight tickets, car-trail for reserving rental cars, and so on (Sen *et al.*, 2006). These trails come to existence because customers are driven by those clear purposes and navigate through the website on more predictable tracks generally consisting of the searching, viewing, booking, and paying steps. However social tagging systems are complex information seeking environments where users are driven by their interests on the fly. Being confronted with the numerous potentially interesting links on every webpage, they are actually given numerous navigation options before every movement rather than guided to a certain direction. So their tracks are too diverse to form trails.

Footprint Level Analysis

As can be told from the *REQ* column of Table *cleaned_data*, users requested 10,303,684 Douban webpages on December 12, 2008, namely, they left 10,303,684 footprints in the website on that day. Going through such a huge number of footprints one by one was neither feasible nor efficient. A better way was reducing them to several principle categories, since each requested page could be categorized, according to its URL, into one of the page clusters mentioned in Section 3.2. Such categorization not only helped us get a general understanding of the scope and center of Douban users' activities, but also laid the ground for the following movement and track level analyses.

Sorting Table *cleaned_data* by *REQ* could group similar URLs together, but it still depended on the researcher to scan the column from top to bottom manually to decide which cluster each group of URLs belongs to.

The result of the initial analysis was the taxonomy of footprints in Douban (Appendix A) with eight essential categories – home, resource, tag, user, group, "mine", search, and collect. If not fitting in with any of them, a footprint belongs to the peripheral category. Some of the essential categories contain one or more subcategories. The footprints in each subcategory are summarized with the page URLs, briefly described, and assigned an abbreviation and a code. Such taxonomy made necessary preparations for the subsequent analysis: the URL keywords were used to construct SQL queries to select the footprints from the tables, and the abbreviations were used in footprint and movement denotation and the codes in track visualization.

In that resource finding and collecting are the focus of research, the analysis at this level subsequently narrowed down to the R and C footprints. They set the milestones on an information seeking path: a R footprint is left when the user accesses a resource page (e.g. "*/subject/3189420/*"); and a C footprint is left when the user performs a resource adding action (e.g. "*/j/subject/3189420/interest?interest=collect&rating=5*"). The 7-digit numbers in the URLs are resource IDs. Douban assigns a unique ID to every resource. For the purpose of providing an overview of the two most important types of user-resource interaction, two new tables, *resource_finding* and *resource_collecting*, were generated with SQL Queries 1 and 2 that respectively select all the resource finding records (R footprint in the *REQ* column) and all the resource collecting records (C footprint in the *REF* column) from Table *cleaned_data*.

UID	RID	REQ	REF	TIME
1017886990	2304115	/subject/2304115/?i=0	/subject_search?cat=1001&search_text=%E6%AF%94%E8%BE%	7:53:24 PM
1961049911	2342570	/subject/2342570/?i=1	/subject_search?cat=1002&search_text=+%E6%9F%B3%E4%BA	7:53:24 PM
1965504750	2228604	/subject/2228604/?rec=1	/movie/	7:53:24 PM
1968229564	1780749	/subject/1780749/?rec=V	/	7:53:24 PM
1968765005	1918707	/subject/1918707/	/subject/1389535/	7:53:24 PM
1969041548	2311147	/subject/2311147/	/subject/2157131/	7:53:24 PM
2005084022	1307657	/subject/1307657/	/movie/tag/%E7%A7%91%E5%B9%BB?start=160	7:53:24 PM
2045420428	1891179	/subject/1891179/	/	7:53:24 PM
2071613577	1863731	/subject/1863731/?i=0	/subject_search?search_text=%E7%A5%9E%E6%8E%A2%E7%A.	7:53:24 PM
2103294700	1305472	/subject/1305472/?i=88	/subject_search?start=75&search_text=%E4%BB%BB%E8%BE%E	7:53:24 PM
2105515094	3322741	/subject/3322741/?i=0	/subject_search?search_text=%E5%A6%82%E6%9E%9C%E4%Bi	7:53:23 PM
-574095283	3048031	/subject/3048031/?i=0	/subject_search?search_text=%E6%B3%AA%E7%97%95%E5%8!	7:53:24 PM
-587168995	1891179	/subject/1891179/	/subject/1891179/discussion?start=60	7:53:24 PM
-588756536	1457449	/subject/1457449/?i=5	/music/search/The%20Seatbelts	7:53:24 PM
-592940630	1482072	/subject/1482072/?i=0	/movie/search/Anne%20Hathaway	7:53:24 PM
-612303642	1424741	/subject/1424741/	/subject/1467776/	7:53:24 PM
-624567828	2170629	/subject/2170629/	/doulist/61053/	7:53:24 PM
-636274061	2007083	/subject/2007083/?i=0	/subject_search?search_text=%E5%8D%97%E6%96%B9%E7%9A	7:53:24 PM
-745952792	1292220	/subject/1292220/	/subject/1292220/edit	7:53:24 PM
-876888155	1295873	/subject/1295873/	/subject/1293234/	7:53:24 PM
975530174	1299059	/subject/1299059/	/subject/1294114/	7:53:24 PM

Figure 6. A snippet from Table *resource_finding*

UID	RID	REQ	REF	TIME
1026613090	1787981	/j/subject/1787981/interest?interest=do	/subject/1787981/	11:10:02 AM
1033415492	1016060	/j/subject/1016060/interest?interest=collect	/subject/1016060/	11:09:59 AM
1124700798	1471556	/j/subject/1471556/interest?interest=wish	/subject/1471556/?rec=1	11:10:02 AM
1950746264	1300299	/j/subject/1300299/interest?interest=wish	/subject/1300299/	11:10:05 AM
1961113862	3238176	/j/subject/3238176/interest?interest=do	/subject/3238176/	11:09:58 AM
2032304833	1048209	/j/subject/1048209/interest?interest=collect&rating=	/subject/1048209/	11:10:04 AM
2073359707	1308807	/j/subject/1308807/interest?interest=collect&rating=	/subject/1308807/?i=0	11:10:04 AM
2085538177	3156578	/j/subject/3156578/interest?interest=collect	/subject/3156578/	11:09:58 AM
-554224332	1422089	/j/subject/1422089/interest?interest=wish	/subject/1422089/	11:09:59 AM
-587635321	2042226	/subject/2042226/?interest=collect&ck=RQe3	/subject/2042226/?i=0	11:09:58 AM
-591470487	3268216	/j/subject/3268216/interest?interest=collect	/subject/3268216/	11:10:03 AM
-635681130	1819912	/j/subject/1819912/interest?interest=collect	/subject/1819912/	11:10:02 AM
-636185912	2059456	/j/subject/2059456/interest?interest=collect	/subject/2059456/	11:10:03 AM
-636185912	1293422	/j/subject/1293422/interest?interest=collect	/subject/1293422/	11:10:06 AM
-636363481	1896550	/j/subject/1896550/interest?interest=collect	/subject/1896550/?rec=A	11:09:59 AM
-637161886	1926728	/j/subject/1926728/interest?interest=wish	/subject/1926728/?i=0	11:10:06 AM
-769610710	1292276	/j/subject/1292276/interest?interest=wish	/subject/1292276/?i=0	11:09:59 AM
974356089	1297102	/j/subject/1297102/interest?interest=collect	/subject/1297102/?from=mb-86815121	11:09:59 AM
989245374	2132495	/j/subject/2132495/interest?interest=collect	/subject/2132495/	11:10:01 AM
993071334	1303394	/j/subject/1303394/interest?interest=collect	/subject/1303394/	11:10:03 AM
994221281	1409704	/j/subject/1409704/interest?interest=collect	/subject/1409704/	11:10:07 AM

Figure 7. A snippet from Table *resource_collecting*

A new field *RID* was added to each table, displaying the resource IDs extracted from the

URLs of the R or C footprints, as seen Figures 6 and 7. By combining the *UID* and *RID* fields, one could reveal a series of facts about the resource finding and collecting occurrences from the tables, including which users viewed/collected resources, how many resources they each viewed/collected (SQL Query 3/4), which resources were viewed/collected, and how many users by whom each resource was viewed/collected (SQL Query 5/6). They were further mapped onto log-log scale plots to test whether the user-resource interaction followed the power Law. A power-law distribution has been detected in query terms submitted to Web search engines (Spink, *et al.*, 2001), as well as in search result clickthroughs and reformulation of queries (Mat-Hassan & Levene, 2005).

Movement Level Analysis

Every movement leaves a footprint. If the footprint left is neither R nor C, the movement is "transitional". It may act as one of the transitions leading towards a "consequential" movement that leaves the R footprint, following which may be a movement leaving the C footprint, called "pivot". Assuming that a user comes across a tag on Douban's book homepage, accesses a book associated with that tag, and adds the book to his or her library at last. The movement from home to tag, contributing to finding the book afterwards, is a transitional movement. In contrast, the consequential movement here, i.e. from tag to resource, is directly responsible for finding the book. Indicating the user's interest in the book, collecting it is the pivot movement in this process.

The movement level analysis, in the first place, concentrated on all the consequential movements made by Douban users, in order to identify their general information seeking strategies for RQ 1. Consequential movements offer more reliable evidences of the ways users look for

resources, compared to transitional ones. In the above example, the strategy for finding that book is browsing by tag, rather than encountering on home. In Table *resource_finding*, each row explains a consequential movement: the *REQ* field tells us which resource page was accessed, i.e. the current footprint, and the *REF* field from where it was accessed, i.e. the footprint left by the preceding movement.

This latter piece of information is very useful for determining the information seeking strategy adopted for that particular occurrence of resource finding, because the URL in the *REF* field will exclusively fall into one footprint category which in turn can be characterized with one of the four strategies as reviewed in Section 2.1.2. For instance, the highlighted row (the first row) in Figure 6a, with a search result page in its *REF* field, features the searching strategy. By examining the entire *REF* column of Table *resource_finding*, one can distinguish the information seeking strategies ever adopted.

While transitional and consequential movements involve actual changes of location in the website, pivot movements are not movements in the real sense because resources are collected right on the resource pages. Indicating the change of status of a resource to a user, the collecting action is a straightforward yet convincing criterion for judging the successfulness of an information seeking process. This distinguishes social tagging systems from such traditional information seeking environments as search engines where whether users are satisfied with the search results after clicking them through is unknown.

There is an analogy between collecting a resource in a social tagging system and purchasing a product from an online retail store because they both suggest positive evaluation of an item.

E-commerce researchers have been using the "conversion rate", the percentage of customers who purchase from the website in all the visitors, to measure the effectiveness of marketing and merchandising efforts (Lee *et al.*, 2001; Ferrini & Mohr, 2008; Booth & Jansen, 2008). This study, similarly, introduces the "find-to-collect rate" to determine the effectiveness of the information seeking strategies for RQ 1.

Let's denote the number of consequential movements featuring a strategy as N_f and the number of pivot movements as the result of that strategy as N_c. For each information seeking strategy identified, its find-to-collect rate $R_{f\text{-}c} = N_c/N_f$. Higher find-to-collect rate means greater effectiveness. This notion is a little different from the conversion rate of an e-commerce website that concerns about which visitors convert into customers while ignoring the quantity of products each customer purchases during a visit. In a social tagging system, a user may find multiple resources with the same strategy, but not necessarily satisfied with and collecting all of them. Therefore resource finding and collecting occurrences should be counted regardless of users. Specifically, N_f were obtained in Table *resource_finding*, and N_c by jointly querying Table *resource_finding* and *resource_collecting* with SQL Query 7.

Although the effectiveness of a strategy is not interpreted in terms of users, it is still helpful to look at strategy adoption from the angle of users since they are responsible for selecting and applying a strategy when looking for a resource. A starfield visualization was created, in virtue of TIBCO Spotfire Professional[27], to reveal how individual users find each resource. The starfield visualization combines simultaneous representation of large numbers of individual data points with

---

[27] http://spotfire.tibco.com/products/spotfire-professional/exploratory-data-analysis.aspx

an interactive interface that allows zooming, filtering, and dynamic querying (Hochheiser & Shneiderman, 1999).

Track Level Analysis

The track level analysis was conducted to answer RQ 3, namely, to disclose how Douban users navigate through the website to the resources they are interested in. Being retrospective, a track takes shape after the user completes all the movements during one website visit. Given that every movement represents a clickstream record, a track can be extracted from the log file by grouping a collection of temporally close clickstream records of a user. This was the first and foremost step to prepare the data for analysis, followed by track normalization.

The extraction of tracks was restricted to regular Douban users, without considering random visitors whose information seeking behavior in this social tagging system was hardly established and thus not worth researching. But we could not tell them from each other directly since the majority of the users did not sign in when visiting Douban. For this reason, a list of active registered users (Table *registerd_uids*) was requested from Douban separately. Their clickstream records were selected into Table *regular_data* from Table *cleaned_data* with SQL Query 8. The researcher then wrote a VBA macro and ran it on the new table after sorting it by *UID* and *TIME*. This macro, named *identify_tracks*, separates two adjacent clickstream records into two tracks when they belong to two users or when the time difference between them exceeds 30 minutes if they belong to one user. The timeout value chosen is common in many websites which automatically end users' sessions after 30 minutes of inactivity.

With the boundaries being drawn between different tracks, it was found that certain clickstream records were absent from each track. Obviously, the first record of a track always lacks a preceding record explaining its *REF* field. And whenever the *REF* field of a non-first record does not match the *REQ* field of any record before it in the same track, there must be a record lost. Such missing data, which might be not available in the original log file or have been cleaned for some reason, was restored after running Macros *restore_first*, *insert_first*, *restore_interrupted*, and *insert_interrupted* on Table *regular_data* in this specific sequence. For each record in the table, if data missing was detected, a restored record would be inserted prior to it. The *UID* field of the restored record was kept the same as the current record, *REQ* field replicated the URL in the *REF* field of the current one, *REF* field was set to null, and *TIME* field was adjusted to be 2 seconds ahead of the current timestamp.

Thanks to the above data preparation, the tracks are ready for analysis on the basis of four general attributes: *length*, *duration*, *capacity*, and *achievement*. The length of a track is the total number of its constituent movements, and the duration is the time difference between the first and last movements. The capacity of a track refers to the quantity of consequential movements on it, and the achievement refers to that of pivot movements. Together they offer a basic description about a visit to the Douban: how many webpages have been accessed during the visit, how much time has been spent on the visit, and how many resources have been viewed and collected during the visit.

The values of the four attributes were obtained for all the tracks respectively through SQL Query 9, Macro *track_duration*, SQL Query 10, and SQL Query 11. A number of zero-capacity tracks were noticed. They represented those visits that accessed no resource pages, namely, they

diverged from information seeking for exclusively comprising peripheral activities. However the rest were focused tracks which appropriately illustrated the information seeking paths on which the users navigated to one or more resources. The researcher then computed the correlation coefficient of each pair of attributes, based on focused tracks only, to investigate how they relate to one another. The strong relationships identified were mapped onto parallel coordinates visualizations again with TIBCO Spotfire Professional. The parallel coordinates is an approach to displaying multivariate data and presenting the relationships among the variables (Inselberg & Dimsdale, 1991). A parallel coordinates visualization looks like a clutter of overlapping lines, but it can be used to search for predominant trends and exceptions (Few, 2006).

As a matter of fact, the track also has a fifth attribute – *path*. It is the arrangement of movements according to a chronological sequence. Researchers of e-commerce websites believe that visualization is the best way to understand customers' click paths because it allows the discovery of browsing patterns hidden in the raw log files (Ting *et al.*, 2004). A fundamental technique for visualizing clickstream data is the tree-like map where the nodes are webpages and they are connected with directed edges (Dömel, 1994; Hirsch *et al.*, 1997; Hong *et al.*, 2001), with the time factor poorly presented. A more recent development by Ting *et al.* (2007) – the footstep graph visualizes the changes of position from one page to another over time in virtue of lines and steps, but these representations are not intuitive enough however.

This study, taking into account their strengths and weaknesses, proposed the "pathway graph". On a 2D polyline graph, the X-axis represents time and Y-axis footprint category (see the code of each category in Appendix A). Each track is made up of nodes and arrows which

respectively stand for footprints and movements. Figure 8 is the pathway graph of a very simple linear track. One can tell from it that the user has performed three searches (code = 7.0) and found two resources (code = 5.0), and whole process lasts about 5 minutes.
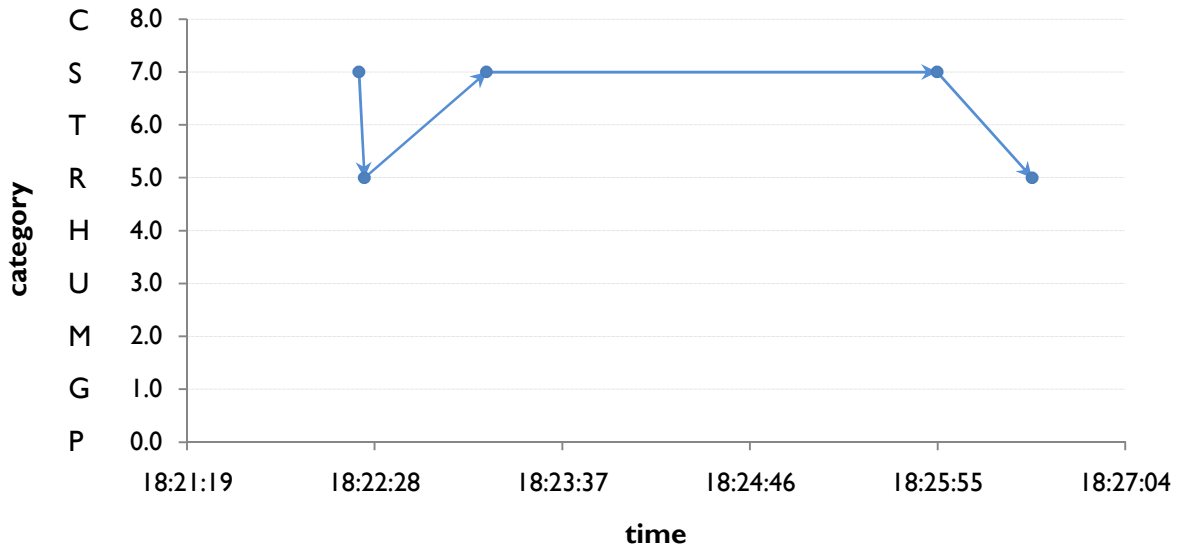
Figure 8. The pathway graph of a linear track

Realistic tracks can be extremely complex. One difficulty in visualizing a track is caused by the fact that users often multitask, i.e. clicking multiple links on a certain page in a row. Reflected in the visualizations are multiple arrows originating from a same node. For example, Table 1 shows four adjacent clickstream records of User *976659235* from Table *regular_data* sorted by *UID* and *TIME*. They describe the multitasking behavior that opens three different resource pages on Douban's music homepage. The first one is the initial record of all the three multitasking records below, but it only appears once. For the convenience of visualization, it should be replicated prior to each multitasking record, which can be achieved with Macros *repliate_multitask* and *insert_multitask*. The replicated records are displayed with bold font in Table 2.

Table 1. Example of multitasking records in Table *regular_data* before replication

UID	REQ	REF	TIME
976659235	/music/	/book/	9:54:24 PM
976659235	/subject/1394547/?rec=1	/music/	9:54:31 PM
976659235	/subject/1455839/?rec=1	/music/	9:55:54 PM
976659235	/subject/1759319/?rec=1	/music/	9:58:59 PM

Table 2. Example of multitasking records in Table regular_data after replication

UID	REQ	REF	TIME
976659235	/music/	/book/	9:54:24 PM
976659235	/subject/1394547/?rec=1	/music/	9:54:31 PM
976659235	**/music/**	**/book/**	**9:54:24 PM**
976659235	/subject/1455839/?rec=1	/music/	9:55:54 PM
976659235	**/music/**	**/book/**	**9:54:24 PM**
976659235	/subject/1759319/?rec=1	/music/	9:58:59 PM

Given the large number of tracks, not all of them were visualized. Double-movement tracks (length = 2) did not deserve visualizing. Indicating casual visits to the website, they are too simple to show any pattern. Therefore, only 15,819 normal and focused tracks (length > 2 and capacity > 0) were considered for visualization. A random sample of 158 tracks (10%) was selected from them, with each sampled track visualized according to its associated clickstream records on a single pathway graph. The visualization was drawn automatically with Macro *visualize_track* in Microsoft Excel in the form of a scatter chart where data points are connected by arrow lines.

3.3.2 Online Survey

Despite that the transaction log provides rich unaltered information about users' behavior, it

contributes little to the exploration of users' personal characteristics which may have direct influences on the ways they behave. It is hence suggested that one should use transaction log analysis in conjunction with other methods to tackle such shortcoming (Jansen, 2008). A natural and excellent complementary data source to the transaction log is the survey, a system for collecting information from or about people to describe, compare, or explain their knowledge, feelings, values, preferences, and behavior (Fink, 2002; Rainie & Jansen, 2008). In the quantitative phase of this dissertation study, particularly, a background survey targeting registered Douban users was conducted as a secondary method to the above clickstream data analysis.

The survey collected data online through a self-administered questionnaire. The questionnaire approach was chosen due to its high efficiency in terms of turnaround time and costs especially when fact-finding questions are adequate for eliciting answers from many different respondents (Fink, 2002). The Web-based style was chosen for the easy access to a fairly large and widely geographically distributed sample (Sue & Ritter, 2007), and more importantly because Douban users are most accessible on the Web from douban.com. A questionnaire form was created in Google Doc and can be viewed here[28]. The language used in the form was Chinese considering that the absolute majority of Douban users are from mainland China. Appendix D includes a translated version of the questionnaire in English.

This online survey featured an unconventional recruitment technique — the "Friend-of-a-Friend" mechanism, which is exclusively feasible in Douban and other websites supporting social networking. At the beginning, the researcher manually selected 10 active Douban

---

[28] https://spreadsheets.google.com/viewform?formkey=dGVSQWRBWDdIbWkzeElDbzZGT3d5ZlE6MQ

users who had built large social networks comprising at least 100 friends, and contacted them via Douban's email service, asking them to fill out the questionnaire form and broadcast the survey invitation to their friends, then friends of friends, and so on. 19 responses were received within the first week. In order to collect more data, more starter users were selected. At the end of the month, the total number of responses reached 129. Not all of them were valid responses and used in data analysis.

The respondents only needed to answer 21 close-ended questions, either multiple choice or checkbox. The questions, as in Appendix D, were clearly organized into three sections asking the respondents to describe (1) their profiles as Douban users, such as their interaction with the system and the details about their libraries; (2) their information seeking experience within Douban; and (3) their basic information as general Web users, including demographics, Web expertise, and search preferences. The analysis of survey data firstly involved running frequency distribution and computing summary statistics for each question. And then the focus was switched to investigating how the user characteristics represented by the questions related to users' adoption of information seeking strategy, by performing chi-square tests. Both the descriptive and inferential statistics were obtained with the help of SPSS.

3.3.3 Focus Group

The second phase of this dissertation study built upon the qualitative method – focus group. "Focus groups are formally organized, structured groups of individuals brought together to discuss a topic

or a series of topics during a specific period of time" (Marczyk *et al.*, 2005, p. 154). Firstly widely used in marketing research, such as on product or program development, customer satisfaction, etc., the focus group has evolved into a principal method of qualitative research in academic settings (Krueger & Casey, 2000). Generally, a focus group recruits 6 to 10 participants who have certain characteristics in common that are relevant to a well-defined purpose, and their interactions are guided by a moderator who sets the stage of discussion with prepared questions (Puchta & Potter, 2004).

The focus group participants react to each other in a lively group conversation, investigating the ways they are both similar to and different from each other. They not only need to verbalize what they think, but also have to justify what they say to the peers, especially those different ones. The researchers thus can seek interpretive insights in their comments. At the same time, preceding comments usually establish the context for the following. Through such exchange, a deeper view of the range of participants' thoughts and experiences will surface (Morgan, 1998). In a mixed methods research design like this one, the significance of the qualitative phase consisted in generating interpretation and providing depth for the quantitative phase. The focus group added to both aspects to the greatest extent in virtue of a unique process of sharing and comparing among the participants.

Specifically speaking, the goal of this focus group study was to collect insiders' in-depth opinions on Douban users' information seeking paths as revealed in the clickstream data analysis and online survey. Consequently, how participants would be selected and investigated was determined after the completion of the quantitative phase. Since the ideal candidates of participants were

experienced Douban users, a recruiting notice was added to the end of the survey questionnaire form, asking if the respondents were willing to voluntarily participate in a follow-up study, the background of which was introduced (Appendix D).

Among the 129 survey respondents, 11 provided their Email addresses, showing their interest in this focus group. However 3 respondents were screened out because their main purposes of visiting Douban were not looking for resources. After the 8 participants were selected, the researcher contacted them about where, when, and how the follow-up study was going to take place. But one of them never responded, so finally the focus group was a relatively small one comprising 7 participants. The major advantage of smaller groups is that each participant will stand a better chance of talking (Morgan & Scannell, 1998). Since the participants were ready to share personal feelings about their daily interactions with a familiar system, they had a high level of involvement and expertise in the discussion and expected more time to fully express their opinions.

Considering the probable wide geographical distribution of participants, the focus group was conducted online, which also cut the costs associated with the traditional focus group research (Edmunds, 1999). Avoiding face-to-face communication, online chatting maintained the anonymity of participants, possibly increasing the openness during the discussions. It was also very efficient that every word they say is automatically recorded by the software and no transcription was needed. However this approach had its inherent disadvantages. For example, slower typists produce fewer words within a given time whereas they should be able to contribute more if speaking. And emotional cues, such as participants' facial reactions and tones of voice, are absent (Edmunds, 1999).

The entire focus group session lasted about 1.5 hours. All the participants and the moderator (i.e. the researcher) signed in to Windows Live messenger, an instant messaging software application, at the designated time. At the beginning of the session, the moderator briefed the participants on the quantitative phase of this dissertation study and stated the importance of this focus group discussion. The session then proceeded based on a pre-developed questioning route composed of 7 questions. The participants were encouraged to frequently refer to Douban website when making comments in order to evoke their memories about certain experience. The language used by all the individuals involved in the session was also Chinese.

As in Appendix E, the questioning route started with an opening question asking the participants to introduce themselves as Douban users. The intents of this opening question were to help everyone feel comfortable and to get them to talk early in the focus group. The second question aimed to focus participants' attention on the research issues in a broad sense by asking about their information seeking experience in Douban generally. It moved the discussion towards the key problems of central interest to this focus group study. The first two questions were not analyzed independently. Obviously, the following 5 key questions represented the 5 primary findings obtained from the clickstream data analysis and survey data analysis. The most discussion time was spent on these questions, and the answers they elicit constituted the essential content for later analysis.

In that the data of the focus group was captured in the form of group chat history, the subsequent analysis was transcript based (Krueger & Casey, 2000). The unabridged chatting history from the session was exported from the instant messaging software and saved in a Word document

which was then imported into ATLAS.ti[29], the powerful qualitative data analysis software, for content analysis. Content analysis is "a research technique for making replicable and valid inferences from texts (or other meaningful matter) to the contexts of their use" (Krippendorff, 2004, p. 18). It is a data-reduction process in which many words of texts are classified into fewer content categories by human coders or by computer (Weber, 1990). ATLAS.ti perfectly served such purposes by helping the researcher read the transcript, extract and group the related comments for each question discussed in the focus group, make notes on them, code similar or contrasting comments, and develop theories about the quantitative results.

3.4 LIMITATIONS

Given the above methodology, there could be several threats to the validity of this dissertation study. Validity refers to the degree to which a study accurately reflects or assesses what it attempts to measure. Researchers are usually concerned with both internal validity and external validity. The former is the "linking power" (LP) of a study, i.e. its capability of permitting "the inference of whether a cause and effect relationship exists", while the latter is the "generalizing power" (GP) pertinent to "inference of generality" of "the causal relationship beyond the study's particular constellation of circumstances" (Krathwohl, 1998, p. 137).

Internal validity was affected by the problems within the study itself. One intrinsic problem

---

[29] http://www.atlasti.com/

of this dissertation study is that the researcher was the only investigator who collects and analyzes data. The researcher is also a regular user of social tagging systems, including Douban. The interpretation of the research findings could be shaped by the researcher's own experience, which sometimes means bias. But the introduction of the focus group to the study helped minimize such bias. The other problem that might reduce the internal validity of this study was the inadequate control over the selection of respondents for the online survey. Although the "Friend-of-a-Friend" approach was very efficient in the search for potential respondents, it led to low response rate and valueless responses.

If a study lacks external validity, the results are not generalizable or transferable to other groups of interest or other context. Being a representative social tagging system, Douban however is a Chinese language website. The subjects investigated in this study were Douban users who were native Chinese speakers. The research findings may have limited applicability to social tagging systems in other languages, but will be applicable to the Chinese-speaking world. The impacts on the external validity of this study also included the relatively short time span (only 1 day) of the transaction log file requested from Douban. Fortunately the considerable size of the data (exceeding 20,000,000 clickstream records) compensated this to some extent. And since the clickstream data analysis focused on registered users who had been visiting the website for a long time and had stabilized their information seeking habits, the temporal factor was less influential.

4.0 RESULTS

This chapter presents the results obtained from analyzing users' information seeking behavior in a representative social tagging system – Douban. Douban users' clickstream data was examined quantitatively at the footprint, movement, and track levels in virtue of such tools as Access, Excel, and Spotfire. The investigation of their background information captured in the online survey, also falling into the quantitative phase, was completed mainly using SPSS. The focus group discussion generated qualitative data, and a content analysis was conducted on the transcript with the help of ATLAS.ti. The major findings deriving from each of these analyses are described and interpreted in terms of the research questions in one of the sections below.

4.1 CLICKSTREAM DATA ANALYSIS RESULTS

4.1.1 Footprint level analysis results

Serving as the foundation of the whole clickstream data analysis, the footprint level analysis was not targeted at any of the research questions. The foremost step of analyzing the footprints left in Douban was categorizing them. The resulting taxonomy of footprints (Appendix A) consists of eight

essential categories and a peripheral category, which has been mentioned in Section 3.3.1.2. While the home, resource, tag, user, and group categories are embodied in the basic website architecture of Douban (Figure 4), the "mine" category contains singed-in users' footprints in their personal homes specially. Each of these categories features a chief subcategory (the first one), accompanied by auxiliary ones if there are any. For example, the chief subcategory of the user category is the U footprint left when a particular user page is accessed. Due to the limited space on the user page, additional pages are needed to display the user's complete collections, contact lists, and so forth. Viewing the additional pages will leave related footprints that belong to the auxiliary subcategories, e.g. Ur, Uf, etc. The other two essential categories, search and collect, are a little different from the above ones for indicating actions performed rather than accesses to existing pages. The S and C footprints respectively derive from searching and collecting actions as triggered by clicking buttons instead of links. At last, all the non-essential footprints, whose role in information seeking is dispensable, are categorized as peripheral.

The next step of the analysis at this level gave special attention to the R and C footprints. Table *resource_finding* includes a total of 1,016,808 resource finding occurrences, involved in which are 139,874 distinct users and 127,759 distinct resources. According to Table *resource_collecting*, the occurrences of resource collecting add up to 239,463, and they involve 38,251 distinct users and 54,675 distinct resources. That is to say, among the 269,658 distinct users who visited Douban on December 12, 2008 (in Table *cleaned_data*), 52% of them accessed at least one resource page, and only 27% of these resource viewers turned into resource collectors who added one or more resources to their libraries. A user may find or collect multiple resources, and a resource may be

found or collected by multiple users. And finding or collecting can happen between the same pair of user and resource more than once.
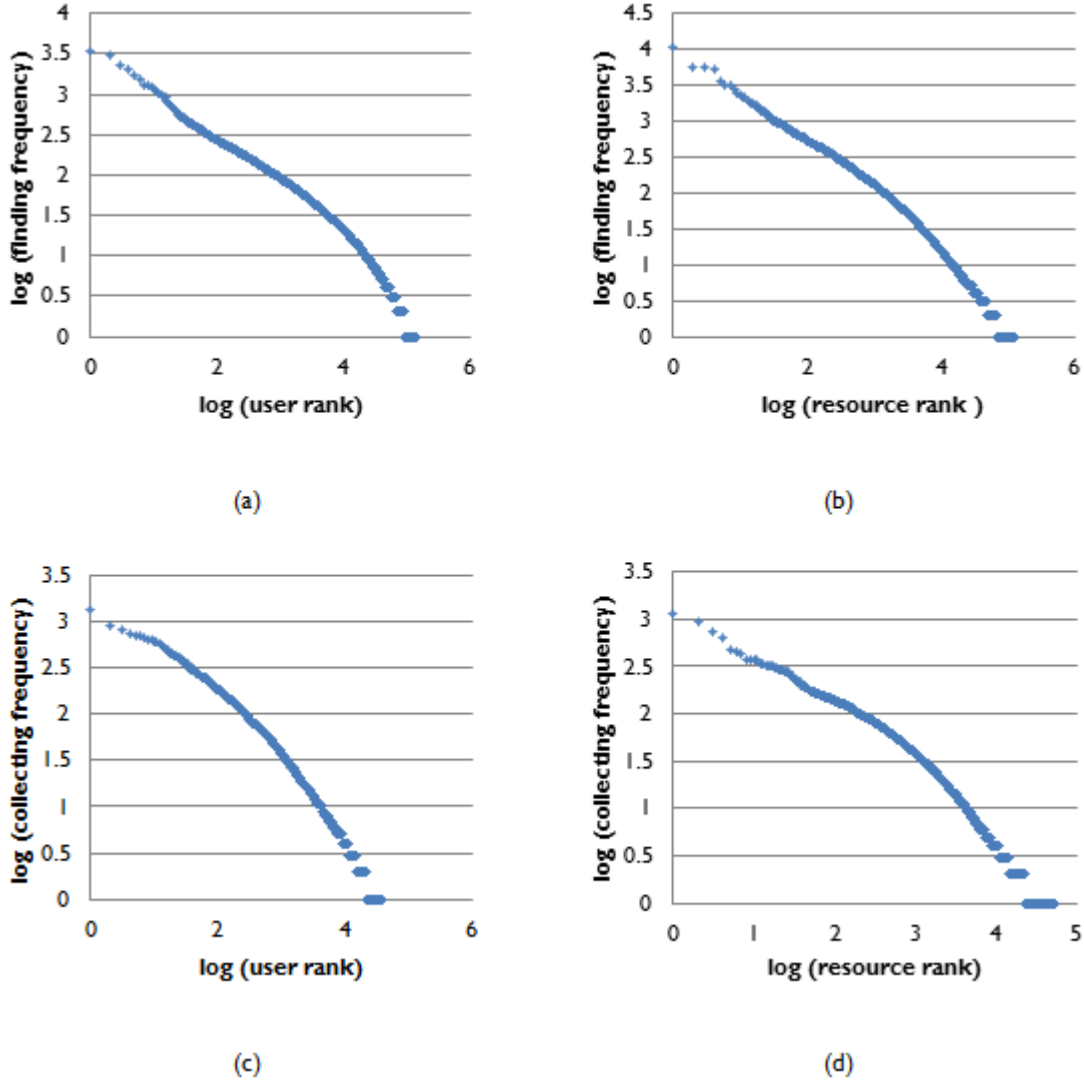
Figure 9. (a) Distribution of finding occurrences among users; (b) distribution of finding occurrences among resources; (c) distribution of collecting occurrences among users; and (d) distribution of collecting occurrences among resources

The distributions of finding/collecting occurrences among users/resources are similar to one another. Each distribution presents a perfect L-shaped curve, with a sharp contrast between the long tail and the top ranked cases. Figures 9a to 9d are the log-log scale plots of the four

distributions. They all show themselves to be linear, which is the characteristic signature of the power law (Adamic, 2000). On the one hand, low activity users or low popularity resources are in the majority. On the other hand, the extreme users or resources involved in exceptionally large numbers of finding or collecting occurrences account for a very small proportion of all the users or resources. 36% of the resource viewers accessed only one resource page, whereas the first ranked viewer accessed 3,431. With 45% of the resources being found only once, the maximum frequency of finding happening to a resource is as high as 10,566. 43% of the resource collectors added only one resource to library, but the first ranked collector added 1,394. While 59% of the resources were collected only once, the maximum frequency of collecting associated with a resource is 1,132.

4.1.2 Movement level analysis results

The movement level analysis addressed RQ 1 in particular. There are three types of movements: transitional, consequential, and pivot. The focuses of analysis were placed on the latter two movements which were systematically investigated to identify the information seeking strategies adopted by Douban users and to determine the effectiveness of each strategy. In Douban, not every page provides links to resource pages. In Table *resource_finding*, a consequential movement can only take one of these 30 forms: H -> R, Hr -> R, Hn -> R, Hp -> R, Hv -> R, He -> R, R -> R, Ru -> R, Rv -> R, Rd -> R, Rl -> R, Rg -> R, V -> R, D -> R, T -> R, U -> R, Ur -> R, Uv -> R, Ue -> R, Un -> R, L -> R, G -> R, Gr -> R, Mr -> R, Fn -> R, Md -> R, Me -> R, Mn -> R, S -> R, and P -> R. Because resources are collected on the resource pages, all the pivot movements in Table

resource_collecting share one uniform notation: R -> C.

However, two sets of consequential movements were not considered in the analysis, including (Ru, Rv, Rd, Rl, Rg, D) -> R and (Mr, Md, Me, Mn) -> R. Ru, Rv, Rd, Rl, and Rg, as auxiliary pages to a resource page, display the resource's complete lists of associated users, reviews, discussion topics, doulists, and groups. They only can be accessed from that specific resource page directly. So if one reaches a resource page from such pages, he or she must have visited the resource page and now is actually returning to it. This is also true to a particular discussion topic page (D) that can be accessed only from R and Rd. Mr, Md, Me, and Mn are exclusively viewable to the current signed-in user. They aggregate the resources which the user has already collected, discussed, recommended, or interacted with somehow, so that the user could revisit those resource pages conveniently. For not reflecting the finding of new resources, these consequential movements were separated from the rest which were then characterized with different information seeking strategies.

Unambiguously, the S -> R consequential movement features the searching strategy. The resource is found as one of the result items on a search result page (S) returned by Douban's internal search engine. The encountering strategy refers to coming across an unexpected resource on the website or sub-site homepages (H, Hr) or other high-level navigational pages with selections of resources (Hn, Hp, Hv, and He).But if one accesses a particular review page (V) from Hv and from there clicks into the corresponding resource page, this is also counted as encountering. The other two strategies, browsing and monitoring, were not readily discernable from each other in the context of Douban. A further examination of the transitional movements prior to certain

consequential movements was needed to find out the strategy.

Assume that we access a resource page from a user page, meaning that the resource may be collected, reviewed, or recommended by the user. If this user is one of our contacts added at an earlier time, it is likely that we have already looked over all his or her resources and the one arousing our attention now is a recent addition to his or her collection. In other words, we are monitoring a known information source, the user, to acquaint ourselves of updates. And Douban offers a shortcut to this end by aggregating our contacts' updates on Fn. In contrast, finding a resource via a newly discovered user is much less directed because we hardly anticipate that some unknown person shares similar interests with us, so the information seeking strategy should be attributed to browsing in this case. In the same way, we monitor the collection of a group we have already joined but browse the collection of a newly discovered group appealing to us, to put it more precisely.

Nevertheless, the browsing strategy takes additional forms. Featured in social tagging systems, browsing the resources associated with a tag (T) is one of them. Douban introduces several loosely defined clusters, such as book clusters "literature", "popular", "culture", "life", "economics", and "technology", to gather similar tags, so that users are able to recognize interesting tags more efficiently when exploring the tag clouds. Another form is browsing the relevant resources of a resource that has been found. The relevance between two resources may be calculated by the system based on the frequency of co-collection or decided by the users based on their own understanding. As a result, one can follow the "people who like this resource also like" (R) or user-compiled doulists (L) on a resource page to find similar resources.

In conclusion, Douban accommodates the following general information seeking strategies which are (1) *encountering*: H -> R, Hr -> R, Hn -> R, Hp -> R, Hv -> R, He -> R, and V -> R; (2) *browsing by resource*: R -> R and L -> R; (3) *browsing by tag*: T -> R; (4) *browsing by user*: U -> R, Ur -> R, Uv -> R, Ue -> R, and Un -> R; (5) *browsing by group*: G -> R and Gr -> R; (6) *searching*: S -> R; (7) *monitoring by user*: U -> R, Ur -> R, Uv -> R, Ue -> R, Un -> R, and Fn -> R; and (8) *monitoring by group*: G -> R and Gr -> R. Indeed, there are other approaches to resource finding in Douban, e.g. discovering a book on the event page about a book signing event. For not being generalizable to other social tagging systems, they were not analyzed.

Table 3. Information seeking strategies adopted by Douban users and their effectiveness

Strategy	N_f	N_c	$R_{f\text{-}c}$
Encountering	163,818	44,548	27.19%
Browsing by resource	327,017	61,109	18.69%
Browsing by tag	113,357	37,756	33.31%
Browsing by user	84,411	13,070	15.48%
Browsing by group	10,330	1,122	10.86%
Searching	264,374	61,919	23.42%
Monitoring by user	18,087	2,456	13.58%
Monitoring by group	1,277	139	10.88%
Total	982,671	222,119	22.60%

The results of the analyses of consequential and pivot movements are shown in Table 3. N_f is the number of consequential movements featuring a strategy, while N_c the number of pivot movements as the result of that strategy. They respectively refer to the frequency of strategy adoption and that of successful strategy adoption. The find-to-collect rate ($R_{f\text{-}c}$) of a strategy is calculated as the ratio of N_c to N_f. The larger the value of N_f, the more popular the strategy is. The

larger the value of $R_{f\text{-}c}$, the more effective the strategy is. The eight general information seeking strategies, as a whole, explain 982,671 resource finding occurrences and 222,119 collecting occurrences in total, and overall find-to-collect rate is 22.60%. However the differences among them are obvious in terms of popularity as well as effectiveness.
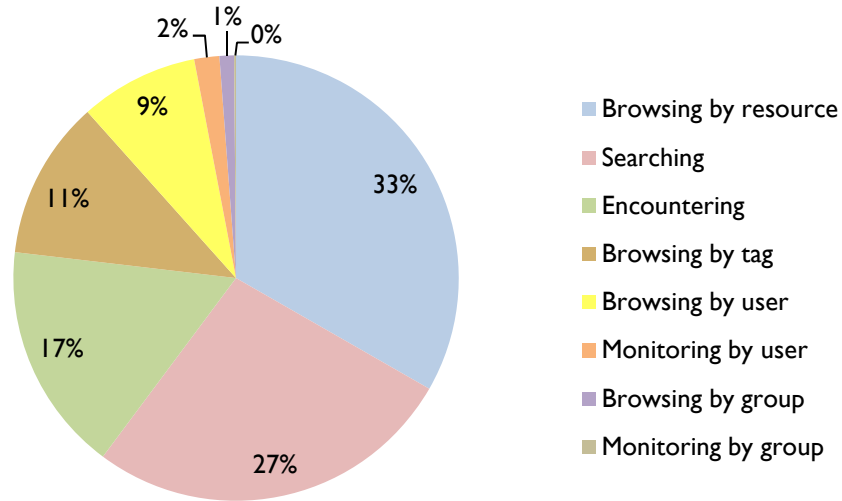
Figure 10. Adoption proportions of different strategies

The most popular strategy among Douban users, as in Figure 10, is browsing by resource. It accounts for almost 1/3 of all the occurrences of resource finding, even 6% more than searching, the most popular strategy among general Web users when they look for information. In respect of browsing relevant resources, users have an obvious preference for co-collected ones over those in the same doulists, with an approximate ratio of 5 to 1. Next to searching is encountering, which is understandable in that a considerable part of what we know is acquired through this passive undirected behavior. The moderate popularity of browsing by tag suggests users' inadequate awareness of the information structure of this social tagging system. The other strategies are all

based on its social structure, involving users or groups as the information sources. By and large, they only play an insignificant or even ignorable role in resource finding in Douban.
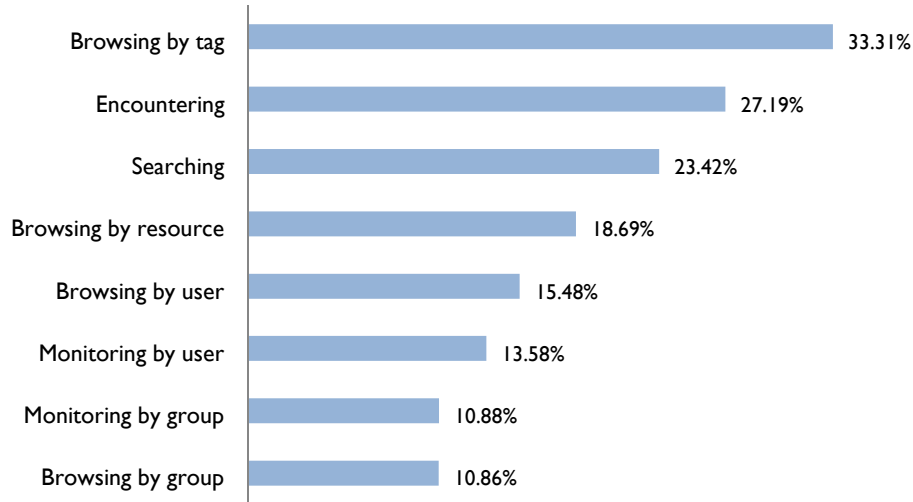
Figure 11. Find-to-collect rates of different strategies

Figure 11 ranks the above strategies according to their find-to-collect rates, from high to low. The chance of acquiring an interesting resource through the exploration of an interesting tag is the highest (33.31%), which makes browsing by tag the most effective strategy, though its absolute frequency of adoption is not competitive. It is a little surprising that the second highest find-to-collect rate (27.19%) is presented by encountering, considering that people cannot anticipate which resources they will encounter. Also in the top three is searching, the purposeful information seeking, which keeps a balance between popularity and effectiveness. Despite the leading popularity of browsing by resource, this strategy helps users find their needed resources only 18.69% of all the times, even lower than the average level (22.38%). The social-oriented strategies, besides less frequently adopted, are also less effective, as can be seen in the bottom half

of Figure 11.

The last objective of the movement level analysis was to visualize the strategy adoption of individual users. The 139,874 distinct resource viewers were ranked according to the number of resources they found. Figure 12 is the starfield visualization of the 5,000[th] ranked user. Each small square in the scatter plot stands for a resource, and the user who found it, time of finding, and strategy used can be told from the X axis, Y axis, and color respectively. This user adopted the searching strategy 12 times, browsing by resource 15 times, and browsing by tag twice, and found 29 resources in total. Theoretically, we can demonstrate all the individual users in a single visualization, with each horizontal line of squares representing one of them. But such visualization has very low legibility due to the huge number of users, which was solved by dividing them into segments.
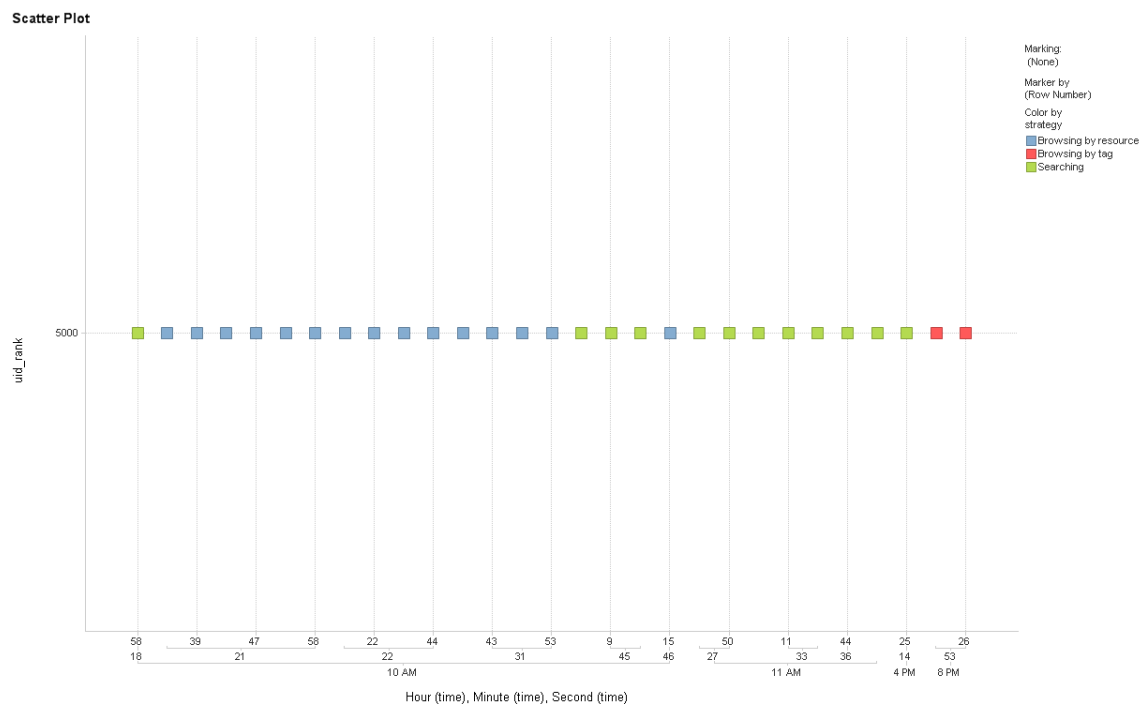

Figure 12. Strategy adoption of the 5,000[th] ranked user

Figure 13 is the starfield visualization of a randomly selected segment of 25 users. Some of the squares on each horizontal line overlap because the user found the resources within a short period of time. This however does not influence the detection of an overt pattern – most horizontal lines are dominated by one color. Namely, most users are accustomed to one specific strategy and adopt it more frequently than any of the other strategies ever adopted. For examples, the 4,050th ranked users exclusively adopted the searching strategy (squares all in green), and the 4,028th ranked user browsing by user (squares all in yellow). As a matter of fact, such domination pattern is ubiquitous in the visualizations of other user segments, implying that the majority of Douban users are describable with their favorite strategies. The preference for a strategy is a habit developed over time. Its establishment can be attributed to many user factors which are however not reflected by the physical behavior. Consequently, this study included an online survey to capture the characteristics of the adopters of different strategies, as will be discussed later.
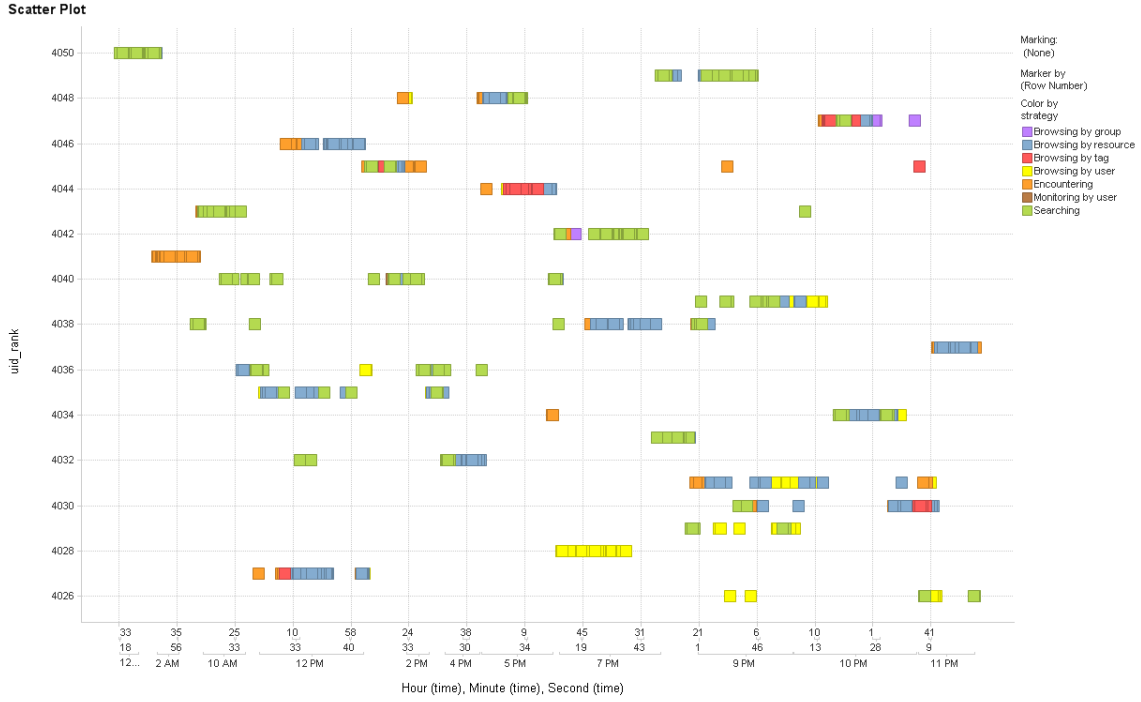
Figure 13. Strategy adoption of the users ranking from the 4026th to the 4050th

4.1.3 Track level analysis results

When visiting a website, the user accesses a series of webpages that are hyperlinked, as represented by movements. A track is composed of all the movements taking place during one visit. The track level analysis aimed at RQ 3. The researcher extracted 37,844 tracks belonging to 8,948 registered users who visited Douban on December 12, 2008, which means an average of 4 visits to this social tagging system per user per day. One-time visitors are in the minority (16.33%), and the highest visiting frequency is 18. If a user has two or more tracks, they do not necessarily resemble each other because the user may behave differently during different visits. So every track was treated independent of the associated user and analyzed in terms of four attributes – length, duration, capacity, and achievement. Table 4 summarizes the descriptive statistics of all the attributes.

Table 4. Descriptive statistics of length, duration, capacity, and achievement

	Length	Duration	Capacity	Achievement
Mean	46.745	31.574	5.049	1.055
Standard Error	1.144	.331	.147	.051
Median	12	9.467	1	0
Mode	2	.033	0	0
Standard Deviation	222.539	64.479	28.683	9.878
Sample Variance	49,523.557	4,157.566	822.718	97.579
Kurtosis	2,906.946	55.698	3,651.379	1,829.427
Skewness	41.732	5.964	46.707	35.568
Range	21,270	1,119.717	2,925	753
Minimum	2	.033	0	0
Maximum	21272	1119.750	2925	753
Sum	1,769,015	1,194,899.683	191,060	39,939
Count	37,844	37,844	37,844	37,844

Length and duration are two basic parameters of the span of a track. The lengths of the 37,844 tracks range very widely, from 2 to 21,272 pages, with a huge standard deviation of 222.539. It can be inferred that the scope within which Douban has been explored is different from visit to visit. The duration of a track is measured in minutes. The large standard deviation (64.479) of track durations suggests considerably high variability among the time spent on the individual visits. While the shortest visits lasted for only 2 seconds, the longest one occupied the user for more than 18 hours. Due to a number of extreme length and duration values, the medians are better measures of the central tendencies of the two attributes than the means. And the median length (12 pages) and median duration (9.467 minutes) do approximate Douban's traffic statistics. According to Alexa[30], during the last quarter of 2008, the average "pageviews" and "time on site" of the website

---

[30] http://www.alexa.com/siteinfo/douban.com

respectively fluctuated around 15 pages and 10 minutes. The modes of durations and lengths, interestingly, both equal the minimum values. This indicates that casual visits occurred in Douban most frequently.

What should be mentioned in particular are a small number of long tracks: 133 tracks are longer than 1,000 pages, and 154 longer than 8 hours. If judging from their extremely large length and/or duration values, one may think that they reflected robots' activities within the system. In fact, however, they resulted from real human activities, as found in the log file. Take Track 29418, which belongs to User 1032861230 who accessed 21,272 pages (first-ranked in length) during one single visit that lasted more than 17.5 hours (second-ranked in duration), for example. As in Figure 14, this is reasonably a human track for comprising clickstream records that were generated in a consecutive sequence at a normal pace. Long tracks represent thorough visits to Douban which are more likely to happen to exploration-oriented users. Such users may not feel the pressure to complete specific tasks, and the abundance of time allows them to examine any information attracting their attention.

Figure 14. A snippet from Table *regular_data*

Finding resources and collecting resources are the two most important types of events during a visit from the perspective of information seeking. The capacities (SD = 28.683) and achievements (SD = 9.878), respectively referring to the numbers of resources found and collected, vary among the tracks to a lesser degree. On average, the users found five resources (Mean = 5.049) and collected one of them (Mean = 1.055) during each visit. The frequency distributions of capacity (Skewness = 46.707; Kurtosis = 3651.379) and achievement (Skewness = 35.568; Kurtosis = 1829.427) values are both seriously skewed to the low end and remarkably peaked at the lowest value (Mode = Minimum = 0). Their shapes are similar to that of the length value distribution (Skewness = 41.732; Kurtosis = 2906.946). In contrast, the duration values demonstrate a much less skewed (Skewness = 5.964) and more dispersed (Kurtosis = 55.698) frequency distribution.

The total number of focused tracks, i.e. capacity > 0, adds up to 21,583. The relationships among the four attributes can be understood through a matrix of correlations created based on all the focused tracks (Table 5). The relationship between length and capacity (r = .889) and that between capacity and achievement (r = .798) are both very strong, and the direct relationship between length and achievement (r = .619) is strong too. It can be inferred that the more the pages accessed during a visit, the more the resource pages accessed, and in turn the more the resources added to library. In contrast, duration is only moderately related to capacity (r = .530) and weakly related to achievement (r = .346), even if it is related to length strongly (r = .673). The duration of a track, therefore, is a less reliable predictor of the numbers of resources found or collected on that track than its length.

Table 5. Correlation matrix of length, duration, capacity, and achievement

	Length	Duration	Capacity	Achievement
Length	I	-	-	-
Duration	.673	I	-	-
Capacity	.889	.530	I	-
Achievement	.619	.346	.798	I

To provide more insights into the ways track length relates to capacity and achievement, their relationships were mapped onto a parallel coordinates visualization (Figure 15). Each polyline in this visualization represents a track, and its length, capacity, and achievement values can be read from the corresponding parallel axes. All the values have been normalized as percentages, with the highest value at the top (100%) and the lowest at the bottom (0%). For example, the track with the longest length, as highlighted in orange, also has the highest capacity but the third highest

achievement. This track and a number of other long-length tracks stand out as exceptions. Their polylines take on various shapes, failing to reflect the strong relationships among the three attributes obtained above.
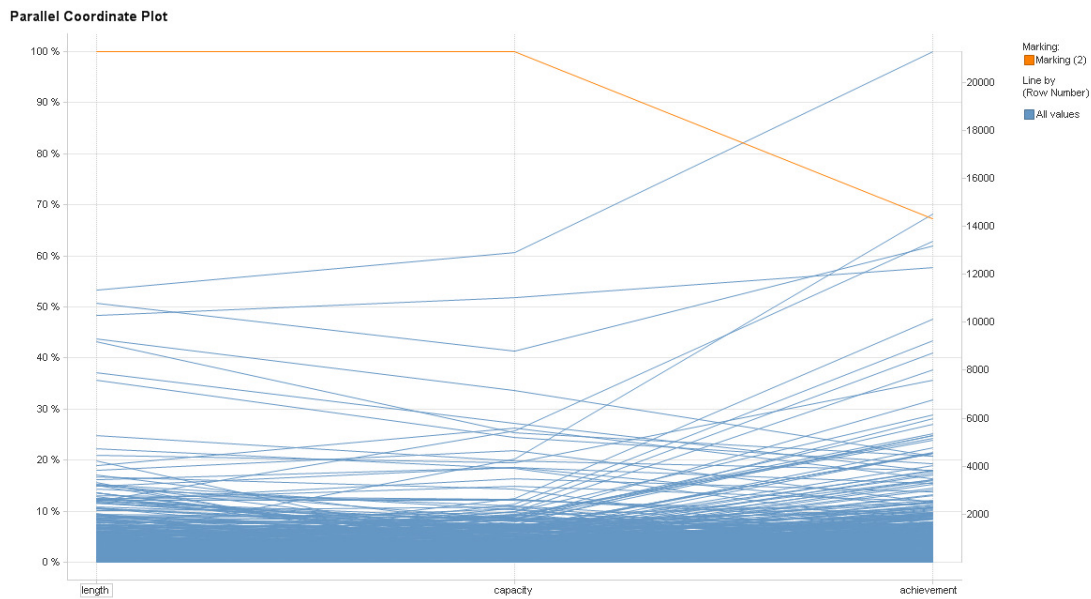
Figure 15. Relationships among length, capacity, and achievement
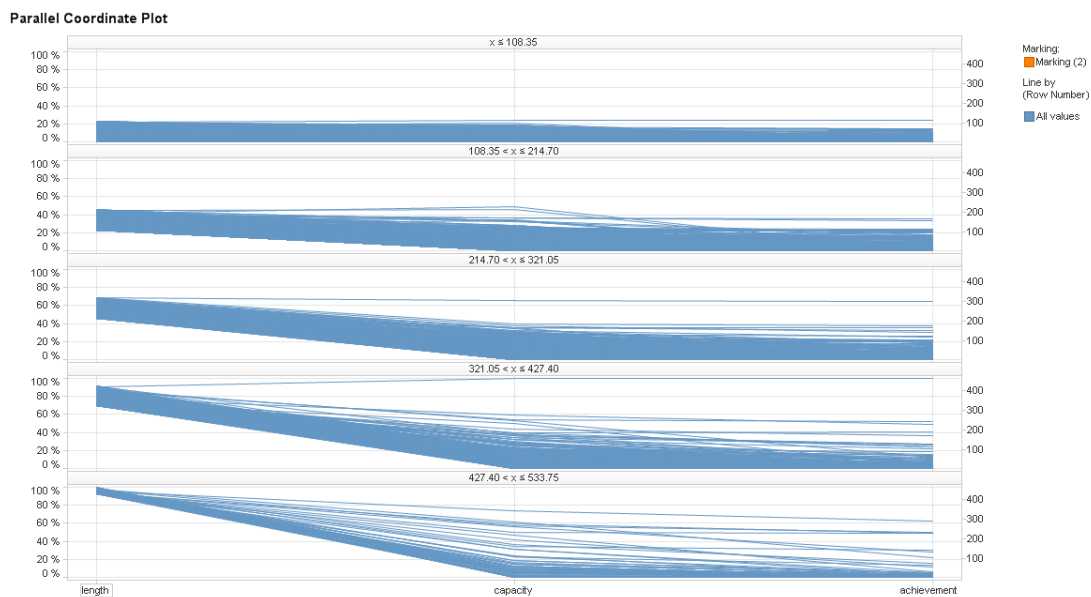
Figure 16. Five groups of short-length tracks

The absolute majority of all the tracks still cluster at the bottom of the visualization. They were further divided into 5 groups based on length and each group was visualized separately in Figure 16. A quick comparison of these groups reveals an interesting pattern: the distributions of capacity and achievement values tend to broaden as the length values increase. In other words, when more pages were accessed during a visit, it becomes harder to predict how many resources were found or collected during that visit.

The final stage of the track level analysis is visualizing individual tracks. All of the 158 pathway graphs generated were inspected by the researcher manually for common information seeking patterns. As mentioned above, no trail forms in social tagging systems, because the tracks differ vastly from one another. When reading a pathway graph, therefore, the big picture is of little significance. The most revelatory aspects are the representations of consequential and pivot movements in the track visualization, which allows us to know under what circumstances resources are found and collected. Specifically, there exist two contrasting resource viewing patterns — continuous and sporadic, and two contrasting resource collecting patterns — lagged and instant.

The continuous viewing pattern refers to the phenomenon that two or more resource pages are accessed one after another from the same source. Among many other tracks presenting such pattern, Track 31339 is a representative one. The user clicked through two results returned by a search, and the two clicks occurred with very short intervals, i.e. 8 seconds. This instance of continuous viewing is highlighted in orange in the pathway graph of the track (Figure 17). However in Figure 18 that illustrates Track 31871, there are also two resources located through the searching strategy, as highlighted, but they resulted from two separate searches. Besides, the finding of the

first resource happened 20 minutes earlier than that of the second one with other activities taking place in between. So this track is characteristic of sporadic resource viewing: one resource page is accessed from one source at a time, independently.
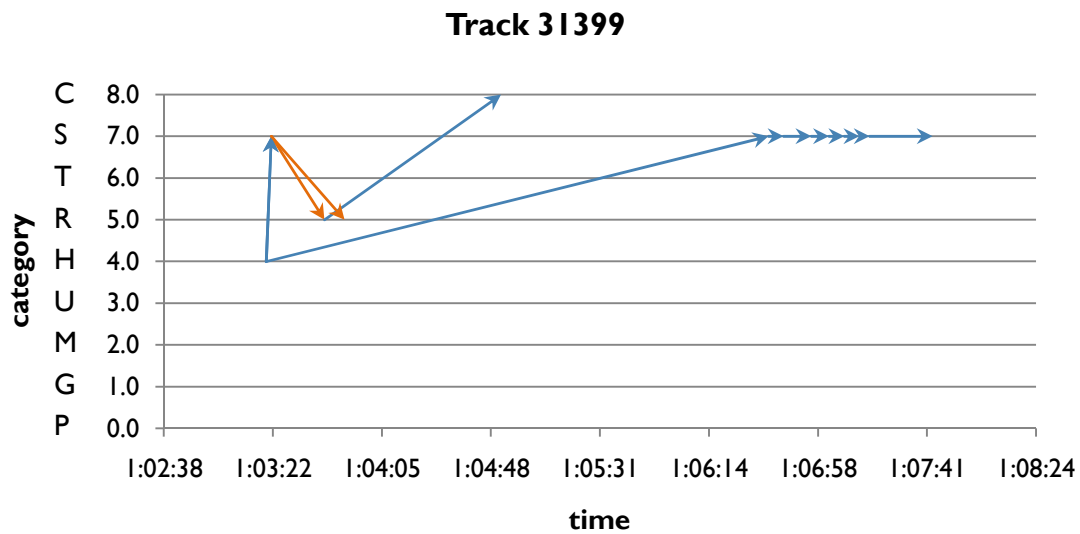
Track 31399

Figure 17. Pathway graph of Track 31399
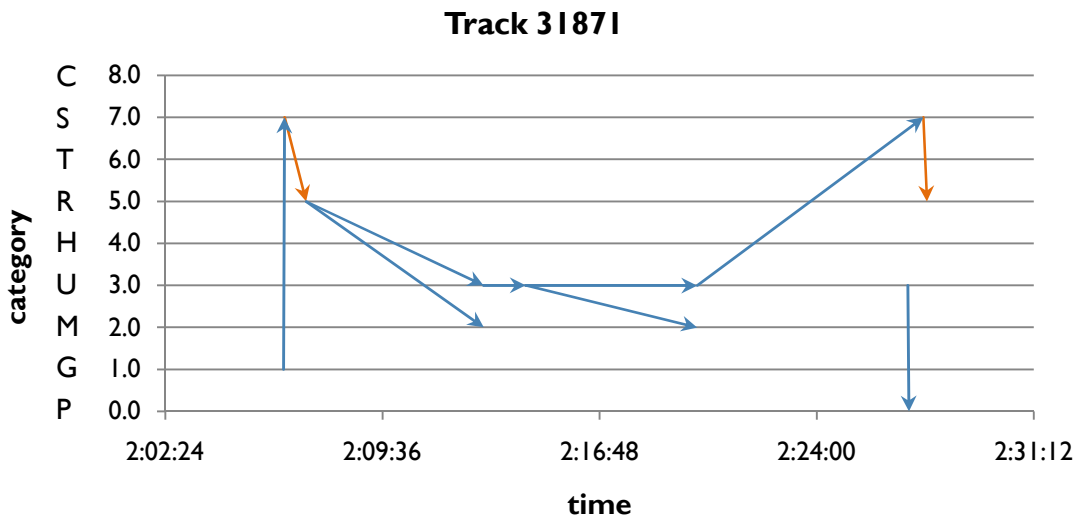
Track 31871

Figure 18. Pathway graph of Track 31871

When it comes to resource collecting, lagged / instant describes how much time the user

spends on a resource page before collecting the resource. In common conditions, there should be a time lag between resource finding and collecting, because in order to decide the interestingness of the resource, the user needs to invest normal time scanning its original information or reading other users' reviews or discussion topics about it. For instance, on Track 13607 (Figure 19), the user showed extra attention when collecting a resource encountered on the homepage. The resource was collected about 8 minutes later after being found, and during that time period the user read a review carefully. Contrary to lagged collecting, instant collecting is preceded by a minimal decision making process. However, on Track 13626 (Figure 20), the resource found on a search result page was collected almost immediately. The user only gave a 3-second glance over the resource page and determined that it was what he or she wanted.
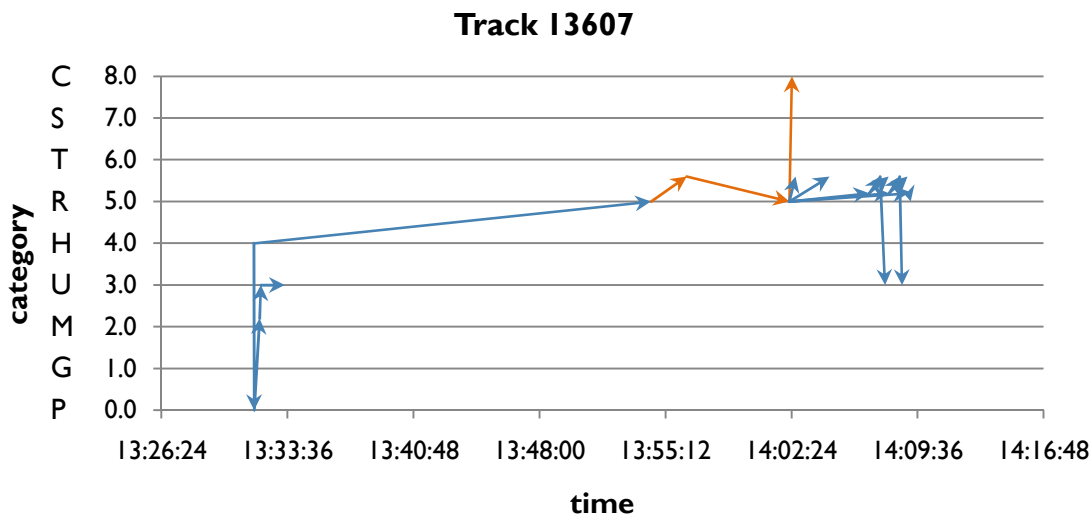
Figure 19. Pathway graph of Track 13607
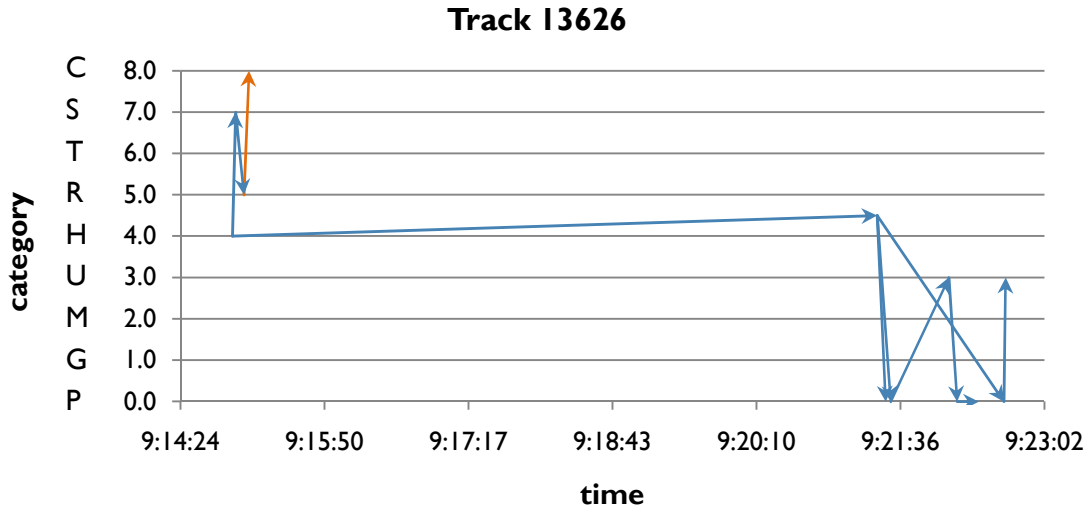
Figure 20. Pathway graph of Track 13626

The continuous and sporadic finding patterns, in fact, are not mutually exclusive from each other, whereas one of them may dominate a track. Meanwhile, the lagged and instant collecting patterns can also co-exist on a track, depending on how the user deals with each resource. One interesting discovery is that the domination of the continuous finding pattern is common on high-capacity tracks and often accompanied by the instant collecting pattern. The convenience of continuous finding consists in that one information seeking effort will lead the user to multiple resources. But when batches of resource pages are accessed on a track, it is not likely that the user can allocate much time examining the details of every one. So if a resource appeals to the user, the collecting tends to be instant. Take Track 3144 for example. In its pathway graph (Figure 21), the highlighted areas represent three information seeking phases largely separate from other activities. During each phase, the resources were found continuously, and all of them were collected, instantly.
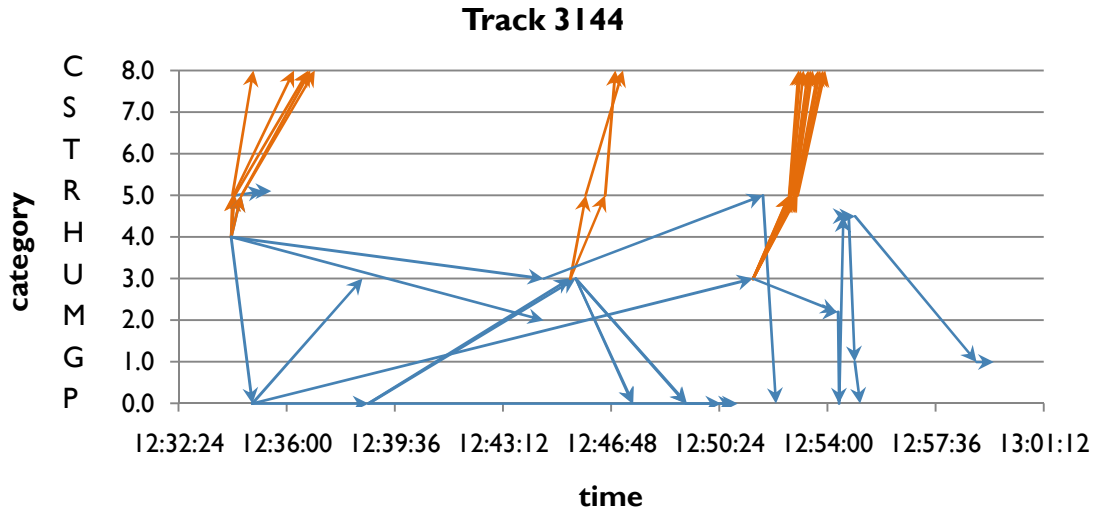
Figure 21. Pathway graph of Track 3144

4.2 SURVEY DATA ANALYSIS RESULTS

4.2.1 Summary of survey data

After the data collection of the online survey was completed, the researcher received a total of 129 responses from registered Douban users who visited the website regularly. Since "Friend-of-a-Friend" mechanism made it impossible to trace how many survey invitations were sent, there was no way to obtain a response rate. Among the 129 collected responses, 17 of them were not valid because the respondents provided conflicting answers. For examples, the quantity of resources collected during a visit (Q 11) exceeded that of resources found (Q 10), and the most frequently used method (Q 13) was not among the previously selected ones (Q 12). They were excluded from the data analysis. The descriptive results were provided below for each question in

the questionnaire.

The 112 respondents' demographic data were collected with Q15 through Q 17. Most of them were in the middle three age ranges 19-22 (N = 42, 37.50%), 23-30 (N = 52, 46.43%), and 31-40 (N = 16, 14.29%). Only 2 were 18 years old or younger, and no one was over 40 years old. Among all the respondents, the females (N = 59, 52.68%) were slightly more than the males (N = 53, 47.32%). Regarding their education levels, 8 respondents completed high school and 4 received other education, while the majority (89.28% in total) obtained higher education degrees, including college (N = 19, 16.96%), bachelor's (N = 40, 35.71%), master's (N =32, 28.57%), and doctoral (N = 9, 8.04%) degrees. As a whole, these respondents reflected Douban's typical users consisting of well-educated young adults of both genders.

Q 18 through Q 21 concern with the respondents' Web experience and information seeking preferences. More than half of them (N = 58, 51.79%) had been acquiring information from the Web for at least 5 years, while short-history (less than 1 year) and medium-history (1 year to less than 5 years) Web users respectively accounted for 12.50% (N = 14) and 35.71% (N = 40) of all. Further, 107 respondents (95.54%) used the Web to look for information at least once a day, with only 5 reporting a frequency of weekly and no one monthly or seldom. So to speak, the respondents were experienced Web users who relied heavily on the Web for information.

Q 20 asked the respondents to enumerate the ways they look for information on the Web. Being one of the four predefined options, "search engines" was selected in 103 responses, "Web directories" 38, "Web portals" 57, and "bookmarked websites" 69. 7 respondents mentioned other methods, including "RSS feeds", "Twitter and micro blogs", and "friends' recommendations", but 2

of them did not specify. 19 respondents (16.96%) were single-method users – 13 of them selected "search engines" solely, 3 "bookmarked websites", 2 "Web portals", and 1 "other". "Web directories" was never unaccompanied. Among the 93 multi-method users (83.04%), 46 combined two methods, 26 three methods, and 20 all four major methods, and one users also used other methods besides the four. In response to Q 21, 62.50% (N = 70) of the respondents indicated that they used search engines most frequently. They far outnumbered the respondents who were accustomed to accessing information via their bookmarked websites (N = 19, 16.96%) and via Web portals (N = 17, 15.18%). Only a few (N =5, 4.46%) gave Web directories first preference. And there was one special respondent whose favorite method was following friends' recommendations.

Table 6 summarizes the respondents' usage of Douban according to their responses to Q 1 through Q 4 as well as Q 10 and Q 11. First of all, they varied greatly in the history and frequency of visiting the website. If we distinguish new and old Douban users with a cutoff history of 6 months, then they each explained about half (49.11% vs. 50.89%) of the survey respondents. However, one can see more frequent users than infrequent ones (65.18% vs. 34.82%), judging from whether they visited the website on a daily basis. This was probably because the former were more accessible by this survey than the latter.

Secondly, the respondents accessed different numbers of webpages and spent different amounts of time per visit, which echoed the high variability among track lengths and durations seen in the track level analysis. Interestingly, for Q 3, the largest proportion (31.25%) of the respondents fell into the ranges "6 ~ 15" that covered the average track length obtained, i.e. 12. But the average track duration, i.e. about 9.5 minutes, belonged to the second most frequently (29.47%) selected

range for Q 4 which was "1 minute ~ less than 10 minutes". More respondents (33.93%) estimated their visit durations to be between 10 and 30 minutes.

Examined in the track level analysis as track capacity and achievement, the quantities of resources found and collected during each visit were also of interest to the survey and captured with Q 10 and Q11. Finding 3 to 5 resources each time was the commonest, with 27.68% of the respondents doing so. The average track capacity obtained, which was 5, remained with this range. The respondents demonstrated less diverse resource collecting habits, comparatively. A considerable proportion (45.54%) of them only added 1 or 2 resources to their libraries each time, very close to the average track achievement which was 1.

Table 6. Survey respondents' Douban usage data

Question	Category	Frequency	Percentage
Q 1	Less than 3 months	21	18.75%
Visiting history	3 months ~ less than 6 months	34	30.36%
	6 months ~ less than 1 year	20	17.86%
	1 year ~ less than 3 years	23	20.53%
	3 years or more	14	12.50%
	Total	112	100.00%
Q 2	More than once a day	36	32.14%
Visiting frequency	Daily	37	33.04%
	Weekly	18	16.07%
	Monthly	8	7.14%
	Seldom	13	11.61%
	Total	112	100.00%
Q 3	≤ 5	17	15.18%
Visit length	6 ~ 15	35	31.25%
	16 ~ 30	25	22.32%
	31 ~ 50	13	11.61%
	> 50	22	19.64%

	Total	112	100.00%
Q 4	Less than 1 minute	22	19.64%
Visit duration	1 minute ~ less than 10 minutes	33	29.47%
	10 minutes ~ less than 30 minutes	38	33.93%
	30 minutes ~ less than 2 hours	11	9.82%
	2 hours or more	8	7.14%
	Total	112	100.00%
Q 10	≤ 2	24	21.43%
Visit capacity	3 ~ 5	31	27.68%
	6 ~ 15	26	23.21%
	16 ~ 30	19	16.96%
	> 30	12	10.72%
	Total	112	100.00%
Q 11	0	20	17.86%
Visit achievement	1 ~ 2	51	45.54%
	3 ~ 5	22	19.64%
	6 ~ 10	14	12.50%
	> 10	5	4.46%
	Total	112	100.00%

Q5 aimed to elicit why the respondents visited Douban. Being consistent with the central role of this social library system, 94 respondents selected "discovering new books, movies, or music albums that I don't know" as the reason or one of the reasons, and 65 respondents selected "collecting books, movies, or music albums that I've heard elsewhere". Both reasons involved resource finding, with the former more like exploratory information seeking and the latter known-item. The two social-oriented reasons, "social networking, i.e. meeting friends, participating in interest groups, etc." and "using other services provided by Douban, e.g. blogs, photo sharing, e-mail, etc.", were selected 51 times and 29 times respectively. Some respondents specified other reasons, such as "book reviews", "music radio service", and "group discussion", which were actually

covered by the predefined options.

In Douban, a user can be profiled from four primary aspects – resources collected, tags assigned, contacts added, and groups joined. Figure 21 shows the respondents' profile data as gained via Q 6 through Q 9. The quantity of resources collected by a user determines the size of his or her Douban library. As seen in Figure 22a, super-small libraries with no more than 10 resources were reported most frequently, by 27 respondents (24.11%), followed by medium (N = 25, 22.32%), super-large (N = 24, 21.43%), large (N = 21, 18.75%), and small (N = 15, 13.39%) libraries. The degree to which a library is organized is reflected by the quantity of tags that the user assigns to his or her collected resources. Figure 22b reveals inadequate efforts devoted to library organization by the respondents. Only 13 (11.61%) of them had well-organized libraries with 101 to 200 tags, and 11 (9.82%) precisely-organized libraries with more than 200 tags. The other respondents' libraries were less organized, including 32 libraries (28.57%) with 51 to 100 tags, 29 (25.89%) with 11 to 50 tags, and 27 (24.11%) with no more than 10 tags.
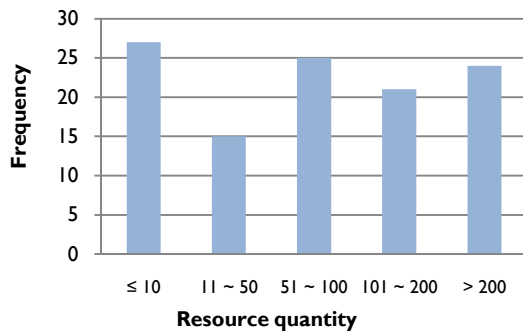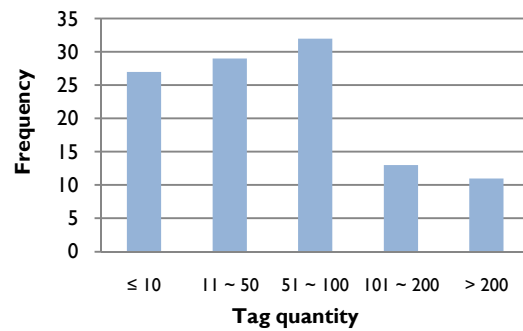
(a)   (b)
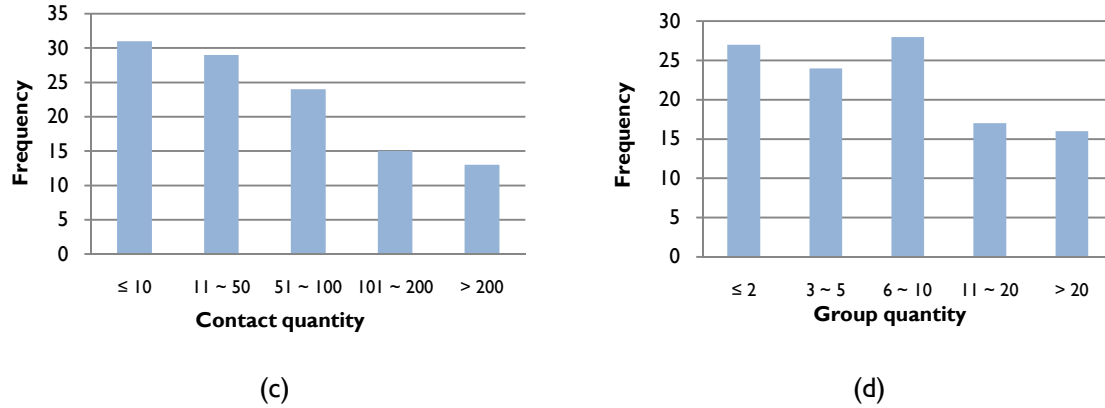
Figure 22. (a) Frequency distribution of resource quantities; (b) frequency distribution of tag quantities; (c) frequency distribution of contact quantities; and (d) frequency distribution of group quantities

Apart from building personal libraries, Douban users also involve themselves in social activities which mainly take two forms: one-to-one interaction with contacts and one-to-many interaction in interest groups. A decreasing trend is obvious in Figure 22c comparing the numbers of respondents in different ranges of contact quantity. 31 respondents (27.68%) added no more than 10 contacts; in contrast, only 13 respondents (11.61%) added more than 200. The higher the range, the fewer the respondents in the range. With regard to group quantity, the decreasing trend is broken by the middle range, as in Figure 22d. The number of the respondents who joined 6 to 10 groups came to 28 (25.00%), which was the largest. The smallest number (N = 16, 14.29%) was associated with the highest range, i.e. ">200". The difference between them was not that big, so the distribution of respondents among the different ranges of group quantity was relatively even.

4.2.2 Exploration of relationships

The movement level analysis has indicated that most Douban users had a favorite information

seeking strategy which could be encountering, browsing by resource, browsing by tag, browsing by user, browsing by group, searching, monitoring by user or monitoring by group. The main purpose of this survey was to help capture the characteristics of different strategy adopters, i.e. tackling RQ 2. As a result, it was necessary to link survey respondents to strategy adopters. Q 13 served such role by asking the respondents to select a method they used most frequently to look for resources in Douban. It was preceded by Q 12 which was interested in their overall strategy adoption. Both questions offered 6 predefined options instead of 8, by merging browsing by user and group into one and merging monitoring by user and group into one.

According to the results of Q 12, "using the internal search engine" was the most popular method, selected in 82 responses. Next to it was "scanning the resources recommended on the homepages" (N = 71), and then "viewing 'people who like this also like' or doulists" (N = 56), "following the tags" (N = 46), "observing the updates in the resource collections of my contacts or affiliated groups" (N = 44), and "exploring the resource collections of random users or groups that I come across" (N = 36). "Other" was selected in 4 responses, one of which specified: "searching Douban with Google". Only 10 (8.93%) respondents were single-method users in the context of Douban: 5 of them selected "using the internal search engine" solely, 2 "following the tags", 2 "scanning the resources recommended on the homepages", and 1 "observing the updates in the resource collections of my contacts or affiliated groups". The other methods were never unaccompanied. Among the 102 (91.07%) multi-method respondents, 22 combined two methods, 49 three methods, 31 four methods and above.

In response to Q 13, no one selected "other". Specifically, 39 (34.82%) respondents selected

"using the internal search engine" as their most frequently used method from all the methods they ever used, 22 (19.64%) "scanning the resources recommended on the homepages", 17 (15.18%) "viewing 'people who like this also like' or doulists", 15 (13.39%) "following the tags", 11 (9.82%) "observing the updates in the resource collections of my contacts or affiliated groups", and 8 (7.14%) "exploring the resource collections of random users or groups that I come across". Respectively, we may call them searchers, encounterers, resource browsers, tag browsers, social monitors, and social browsers.

At last, Q 14 was interested in the respondents' overall information seeking experience. Most (93.75%) of the responses converged at the middle three frequencies. 23 (20.54%) respondents thought that the circumstance described in the question seldom happened to them, 47 (41.96%) occasionally, and 35 (31.25%) frequently. In contrast, only 2 (1.79%) respondents selected "never" and 5 (4.46%) selected "constantly". These frequencies implied the respondents' satisfaction with their information seeking experience within Douban in an opposite direction. Considering that occasional unsuccessful resource finding is acceptable, we may refer to high satisfaction with the first three frequencies and low satisfaction with the last two.

As an important part of the survey data analysis, a series of Pearson's chi-square tests for independence were conducted with SPSS to see if the adopters of the same strategy shared similar demographics, Web experience, search preference, Douban usage, user profile, and satisfaction level. So the column variable in all the tests was strategy adoption in Douban which contained 6 categories as mentioned in Q 13. One of the assumptions of chi-square test is that at least 80% of the cells in the cross-tabulation table have expected frequencies of 5 or more (Pallant, 2007). In

order to decrease the chance of violating this assumption, the researcher reduced the numbers of categories in most row variables to a minimum of two through merging.

Demographic variables include age (younger: 22 years old or younger; and older: 23 years old or older), gender (female; and male), and education level (lower: bachelor's degree or lower; and higher: master's or doctoral degree). For age, χ^2 (5, N =112) = 2.449, p = .784; for gender, χ^2 (5, N =112) = 1.139, p = .951; and for education level, χ^2 (5, N =112) = 5.070, p = .407. It is noticed that all the three p values are larger than the alpha value of .05, meaning that the results are not significant. We can conclude that the respondents adopting different strategies did not have significantly different demographics.

As a general Web user, every respondent provided his or her Web using history (short: less than 5 years; and long: 5 years or over) and frequency (low: less frequently than once a day; and high: at least once a day). For Web using history, χ^2 (5, N =112) = 3.962, p = .555. Namely, there is no significant association between strategy adoption and how long the adopters have been using the Web to look for information. For Web using frequency, χ^2 (5, N =112) = 4.529, p = .476. However 50% of the cells in this test had expected count less than 5, considering that only 6 respondents (5.36%) did not use the Web on a daily basis. Since the above assumption is violated, the result may not be meaningful.

Assumption violation happened again in the chi-square test of the relationship between strategy adoption in Douban and that on the Web (search engines; Web directories; Web portal; and bookmarked websites). Even if the number of categories in the former variable was reduced to four by merging all the browsing strategies together, there were still 8 cells (50%) in this 4 by 4 table

having expected count less than 5. This should be attributed to the overwhelming prevalence of the searching strategy on the Web. As found for Q 21, 62.50% of the respondents selected search engines as their most frequently used method of Web information seeking.

Two basic variables concerning the usage of Douban were the visiting history (short: less than 6 months; and long: 6 months or over) and frequency (low: less frequently than once a day; and high: at least once a day). For Douban visiting history, χ^2 (5, N =112) = 12.595, p = .027; and for Douban visiting frequency, χ^2 (5, N =112) = 4.367, p = .498. Obviously, different strategy adopters did not visit the website at significantly different frequencies. But they differed significantly with regard to how long they have been Douban users, since the former p value is smaller than the alpha value of .05.

Table 7 displays the cross-tabulation between strategy adoption and Douban visiting history. The adjusted residuals were calculated in each cell. According to Aspelmeier *et al.* (2009), if the adjusted residual is greater than or equal to 1.96, then the observed frequency is significantly different from the expected frequency for that cell. Significant adjusted residuals only appear in the two cells of encounterer (z = -2.9, z = 2.9) which contribute significantly to the overall chi-square statistic. It can be inferred that the respondents adopting the encountering strategy tended to be new users, i.e. having short visiting history. As for other strategies, as many adopters as expected were new users and old users.

Table 7. Strategy_adopter * Douban_visting_history Cross-tabulation

	Douban_visting_history		Total
	Long	Short	

Strategy_ adopter	Encounterer	Count	5	17	22
		Expected Count	11.2	10.8	22.0
		% within Strategy_adopter	22.7%	77.3%	100.0%
		Adjusted Residual	-2.9	2.9	
	Resource browser	Count	11	6	17
		Expected Count	8.7	8.3	17.0
		% within Strategy_adopter	64.7%	35.3%	100.0%
		Adjusted Residual	1.2	-1.2	
	Searcher	Count	20	19	39
		Expected Count	19.8	19.2	39.0
		% within Strategy_adopter	51.3%	48.7%	100.0%
		Adjusted Residual	.1	-.1	
	Social browser	Count	3	5	8
		Expected Count	4.1	3.9	8.0
		% within Strategy_adopter	37.5%	62.5%	100.0%
		Adjusted Residual	-.8	.8	
	Social monitor	Count	7	4	11
		Expected Count	5.6	5.4	11.0
		% within Strategy_adopter	63.6%	36.4%	100.0%
		Adjusted Residual	.9	-.9	
	Tag browser	Count	11	4	15
		Expected Count	7.6	7.4	15.0
		% within Strategy_adopter	73.3%	26.7%	100.0%
		Adjusted Residual	1.9	-1.9	
Total		Count	57	55	112
		Expected Count	57.0	55.0	112.0
		% within Strategy_adopter	50.9%	49.1%	100.0%

Length (short: 15 pages or less; and long: more than 15), duration (short: 10 minutes or less; and long: more than 10 minutes), capacity (low: 5 resources or less; and high: more than 5 resources), and achievement (low: 2 resources or less; and high: more than 2) are Douban usage variables about individual visits. For length, χ^2 (5, N =112) = 6.622, p = .250; for duration, χ^2 (5, N =112) = 2.539, p = .771; for capacity, χ^2 (5, N =112) = 15.155, p = .010; and for achievement, χ^2 (5, N =112) = 5.348, p = .375. Except for capacity, the other three results, i.e. the differences

in the numbers of webpages accessed, amounts of time spent, and numbers of resources collected during each visit, are not significant.

Specially, different strategy adopters demonstrated significant differences in the numbers of resources they found during each visit. In Table 8 which is the cross-tabulation table of strategy adoption and visit capacity, significant adjusted residuals appear in four cells, two belonging to resource browser ($z = 2.4$, $z = -2.4$) and the other two belonging to searcher ($z = -3.0$, $z = 3.0$). These tell us that resource browsers tended to find more than 5 resources during each visit while searchers 5 at most. The numbers of resources found by other strategy adopters however were not significantly different than expected.

Table 8. Strategy_adopter *Visit_capacity Cross-tabulation

			Visit_capacity		
			High	Low	Total
Strategy_ adopter	Encounterer	Count	14	8	22
		Expected Count	11.0	11.0	22.0
		% within Strategy_adopter	63.6%	36.4%	100.0%
		Adjusted Residual	1.4	-1.4	
	Resource browser	Count	13	4	17
		Expected Count	8.5	8.5	17.0
		% within Strategy_adopter	76.5%	23.5%	100.0%
		Adjusted Residual	2.4	-2.4	
	Searcher	Count	12	27	39
		Expected Count	19.5	19.5	39.0
		% within Strategy_adopter	30.8%	69.2%	100.0%
		Adjusted Residual	-3.0	3.0	
	Social browser	Count	3	5	8
		Expected Count	4.0	4.0	8.0
		% within Strategy_adopter	37.5%	62.5%	100.0%

		Adjusted Residual	-.7	.7	
	Social monitor	Count	4	7	11
		Expected Count	5.5	5.5	11.0
		% within Strategy_adopter	36.4%	63.6%	100.0%
		Adjusted Residual	-1.0	1.0	
	Tag browser	Count	10	5	15
		Expected Count	7.5	7.5	15.0
		% within Strategy_adopter	66.7%	33.3%	100.0%
		Adjusted Residual	1.4	-1.4	
Total		Count	56	56	112
		Expected Count	56.0	56.0	112.0
		% within Strategy_adopter	50.0%	50.0%	100.0%

The four user profile variables are resource quantity (small: 50 or less; and large: more than 50), tag quantity (small: 50 or less; and large: more than 50), contact quantity (small: 50 or less; and large: more than 50), and group quantity (small: 5 or less; and large: more than 5), respectively. For resource quantity, χ^2 (5, N =112) = 6.641, p = .249; for tag quantity, χ^2 (5, N =112) = 11.230, p = .047; for contact quantity, χ^2 (5, N =112) = 3.226, p = .665; and for group quantity, χ^2 (5, N =112) = 2.005, p = .848. Again, three of the variables fail to show a significant relationship with strategy adoption. The respondents adopting different strategies were not significantly different in the numbers of their resources, contacts, or groups.

Significant differences are seen in the numbers of tags assigned by different strategy adopters. Based on the cross-tabulation table of strategy adopter and tag quantity (Table 9), tag browsers tended to assign large numbers of tags, i.e. more than 50, with significant adjusted residuals appearing in its two cells (z = 2.6, z = -2.6). The numbers of tags assigned by the adopters of other strategies, however, were neither significantly larger nor smaller than the expected values.

Table 9. Strategy_adopter * Tag_quantity Cross-tabulation

			Tag_quantity		Total
			Large	Small	
Strategy_adopter	Encounterer	Count	12	10	22
		Expected Count	10.8	11.2	22.0
		% within Strategy_adopter	54.5%	45.5%	100.0%
		Adjusted Residual	.6	-.6	
	Resource browser	Count	5	12	17
		Expected Count	8.3	8.7	17.0
		% within Strategy_adopter	29.4%	70.6%	100.0%
		Adjusted Residual	-1.8	1.8	
	Searcher	Count	20	19	39
		Expected Count	19.2	19.8	39.0
		% within Strategy_adopter	51.3%	48.7%	100.0%
		Adjusted Residual	.3	-.3	
	Social browser	Count	3	5	8
		Expected Count	3.9	4.1	8.0
		% within Strategy_adopter	37.5%	62.5%	100.0%
		Adjusted Residual	-.7	.7	
	Social monitor	Count	3	8	11
		Expected Count	5.4	5.6	11.0
		% within Strategy_adopter	27.3%	72.7%	100.0%
		Adjusted Residual	-1.5	1.5	
	Tag browser	Count	12	3	15
		Expected Count	7.4	7.6	15.0
		% within Strategy_adopter	80.0%	20.0%	100.0%
		Adjusted Residual	2.6	-2.6	
Total		Count	55	57	112
		Expected Count	55.0	57.0	112.0
		% within Strategy_adopter	49.1%	50.9%	100.0%

The last relationship examined was that between strategy adoption and satisfaction level. If

the resources that the respondents find in Douban are never, seldom, or occasionally not worth

collecting, they will be highly satisfied. But if the worthlessness situation happens frequently or even constantly, their satisfaction level will be low. The result of this chi-square test was χ^2 (5, N =112) = 8.75, p = .121. With a p value that is not significant, strategy adoption and satisfaction level are independent of each other. Put another way, whether a respondent was satisfied with his or her information seeking experience in Douban was not significantly related to what strategy he or she adopted most frequently.

4.3 FOCUS GROUP ANALYSIS RESULTS

During the 1.5-hour discussion, the 7 participants in the focus group generated a total of 341 pertinent messages in Windows Live Messenger. It was common that the participants expressed a complete comment on an issue with several messages. For instance, "*I don't think so*", "*what is the point of including so much content*", and "*users seldom click it through*" were three continuous messages from the same participant who argued against providing a variety of content on the webpages. After merging adjacent relevant messages, the focus group transcript comprised 157 comments. The numbers of comments made by a participant ranged from 12 to 34, with an average of 22. Quantity certainly was not the most important criterion for measuring the contribution of an individual. This content analysis paid a lot of attention to the amount and especially the pertinence of information contained in the comments to decide if they were helpful in understanding the phenomena discussed in the key questions.

4.3.1 Discussion about the popularity of different strategies

Q 3 was the first key question in the focus group and attracted 33 comments. Searching, browsing by resource, and encountering were often mentioned as frequently used strategies, whereas the other three strategies, including browsing by tag, monitoring by user/group, and browsing by user/group, were thought to be less frequently used. The participants offered various explanations for the popularity of different strategies.

The searching strategy was considered from the user, system, and task aspects. It was believed that users' prior familiarity with Web search engines and the great accessibility of the search function in the website both made Douban's internal search engine a necessity to the users. For examples:

◆ "*Everyone uses search engines on the Web… it's a habit*";

◆ "*You can search on any page whenever you need*".

But there was one participant also taking into account the impact of the task. He or she gave an example of the context in which searching was most appropriate – when you knew exactly what you need:

◆ "*If you want to find other people's reviews of Eason Chan's new album, searching is the fastest way*".

The participants deemed browsing by resource and encountering handy strategies because they both enable users to find resources with very low investment in time or effort:

- "*You don't have to look for resources… they will come to you*";

- "*The resources are just there and it doesn't hurt to take a glance*".

In particular, browsing by resource was "*a quick way to look for related resources*". But the adoption of this strategy could be driven by certain user factors, such as curiosity and aimlessness:

- "*Just cannot help clicking on the similar resources when they are at my hand… it's very likely that they will be interesting to me*";

- "*I often look at people-also-like resources when I am bored… just for killing time*".

Encountering resources, which happened on the homepages, was always "*inevitable*". This was because viewing homepages had been integrated into their visiting of Douban by many users:

- "*The homepage is where I start to visit the website every time… if something interesting is there, I will definitely notice*";

- "*Me too… and I think most other users too*".

Among the comments relating to browsing by tag, there were two in-depth ones that explained the less frequent use of tags for resource finding in Douban. One of them considered general Web users' unfamiliarity with tagging and this specific way to look for resources:

- "*The tag cloud is a new Web technology… although many people have a lot of experience with search engines, they are not familiar with tags… as far as I am concerned, I never tagged before visiting Douban*

and at first I would not tag the resources when collecting them… it took me some time to figure out how

tagging worked and what tag clouds were used for";

However, the other comment pointed out a unique problem to Douban users, which was the low usability of the first-generation tag clouds used in the website until April 2009:

◆ "*I want to add… the first-generation tag cloud in Douban was a big mess… it displayed millions of tags on thousands of pages… any tag might appear there, even the ones you couldn't understand… I often avoided using the tag cloud at that time, but I like the new one that is more organized*".

Comparatively, monitoring and browsing by user/group aroused much less discussion among the participants. The only comment referring to the monitoring strategy was:

◆ "*I use this method from time to time… I guess the precondition is that your contacts update their collections frequently*".

Put another way, it had limited applicability. The participants mentioned the same problem about browsing by user/group. For examples:

◆ "*If you ask me, I will not say it is a method for you to look for resources… to me it's more like a part of the process of getting to know a user or a group*";

◆ "*Of course it is… but not working for users who are not interested in discovering people or groups*".

4.3.2 Discussion about the effectiveness of different strategies

The second key question, Q 4, guided the participants to continue their discussion around the strategies, with a focus on their effectiveness in the form of find-to-click rate. Although this question elicited 36 comments, which was the most, none of them provided useful information about monitoring or browsing by user/group. As for the other four more popular strategies, the participants basically agreed on the low effectiveness of browsing by resource, but they had debates on that of the rest in varying degrees.

6 out of 7 participants shared their experience of browsing "people who like this also like", and they mostly complained about the inability of the system to present new resources to users who heavily relied on it or had very large libraries. Here are two examples:

- *"If you view 3 or 4 books continuously in this way, you will find that their people-also-like movies are almost the same"*;

- *"I used this method a lot when I was a new user… it worked well then… but now I have collected more than 700 movies… it is not likely it will bring me new ones… for example, people who like 'Shrek' 4 also like 'Shrek' 1 -3 which I have already collected"*.

A couple of participants thought that the above reason was just secondary, and they had doubt about the relevance between co-collected resources:

- *"Many people-also-like resources are not related in the way you expect… for example, I like love movies*

such as 'Titanic' but not man movies such as 'The God Father'… however other people like both probably because they are classic American movies".

In contrast, the participants conveyed very different opinions concerning browsing by tag, especially the dependability and usefulness of this strategy in the context of Douban. For instance, an opponent of tagging said:

- "*It's little surprising that you can always find what you need with tags… many websites have tag clouds, but I seldom use them… I've read articles talking about the unreliability of user tags*";

This comment was refuted by other participants with confidence in the quality of the tags created by Douban users:

- "*Obviously you were preoccupied by the pessimistic assessment… this is not the case in Douban… here the tag clouds only include the most frequently used tags, meaning they have been agreed by many users*";
- "*Taggers are responsible…irresponsible people will be too lazy to tag… plus, you don't need to come up with a tag yourself because Douban will tell you which tags most people used for a book*".

The precision of the results obtained through browsing by tag was also in controversy. For examples:

- "*I have to say that tags are better than search… it is a difference between human and computer… when I searched for black humor movies, I only got results having the keyword 'black humor' in the titles… but the tag 'black humor' led me to movies in this special comedy form… these are what I need*";

◆ "*Please notice… what you need is not always that general… when you want to find black humor movies in Chinese… Douban does not give you the option to combine the two tags*".

Further, the coverage of this strategy could be problematic to some users who had unpopular interests:

◆ "*What if a resource is never tagged… you will never find it via tags*".

It seemed that these participants were not only used to encountering resources on Douban's homepages, but loved the ones they encountered. This mainly could be understood with their trust in the recommendation system:

◆ "*Douban does a good job in aggregating new resources, good resources, and hot resources, all on its homepages*".

However there were also untraditional reasons, such as:

◆ "*The music resources recommended on the homepage may not fit into my taste… but they are the trend… I don't want to fall behind the times*";

◆ "*Douban won't recommend something too bad… just add them… you can delete the ones you don't like later anyway*".

That is to say, it was possible that users would collect resources not matching their interests; instead, they would develop new interests based on what most people liked.

The searching strategy was mentioned by every participant, and most of their comments were neutral. Namely, whether Douban's internal search engine was effective depended on what kind of task it was used for. Only one participant expressed strong dislike of this search engine at the very beginning and specified its disadvantages as follows:

- *"Like I said, Douban search is useless… what if I don't know the book title or author… it does not accept natural language… for example, when I type 'good books for teenagers' in the search box, no result will be returned"*;

- *"I also have to mention the meaningless ranking of search results… they are not ranked according to relevance, time, rating, or any other specific criteria… this is a big problem when there are too many results"*.

Other participants responded to such negative comments by justifying the incompetence of this search engine. For examples:

- *"Don't be too critical… you cannot expect Douban search to work like Google… given that there are so many ways to look for resources within the website, searching is sort of auxiliary"*;

- *"Ranking is a problem to me too… but I feel OK about Douban search in general… use it when you want to find a specific book"*.

4.3.3 Discussion about the characteristics of different strategy adopters

Q 5 asked the focus group participants to characterize the users who had different favorite strategies, which had been statistically done in the online survey. In spite that the question provided a whole range of characteristics variables which could be considered, not all of them were actually touched in the 25 comments under this key question. Interestingly, the participants mentioned the relationships found significant in the survey data analysis. But they also gave attention to additional relationships that should or should never be valid.

Most participants believed that the way people look for information on the Web could not predict the way they look for resources in Douban because the majority of Web users were searchers. However, one participant also expressed his or her viewpoint about who favored bookmarked websites over search engines:

◆ *"They should like to observe people or groups in Douban… they should have more trust in the resources from somewhere they already know"*.

It was shared by two participants that users who had rich information seeking experience on the Web might have a tendency to look for resources by following tags in Douban. One of them said:

◆ *"I'm not sure if this is true… but at least to me it is… I am a skilled Web user due to my work… I have known tagging since it appeared in Delicious… when I found Douban also supported tagging, I naturally went for tags"*.

It appeared that which strategy a user would adopt most frequently was closely related to his or her visiting history in Douban. The participants thought that new users had a greater chance of preferring to encounter or browse by resource. For examples:

- *"New users were not familiar with the system yet… taking recommendations is the only method that does not require them to be…"*

- *"If you only have a few resources in your library, why not try their people-also-like resources… you can easily increase the size of your collection in this way "*;

What is more, these new users had little chance of being social monitors or tag browsers, because:

- *"Some methods are only applicable to users who have been visiting the website for a while… for example, it takes time for users to build their friend circles… or they have few contacts to observe… and getting to know what tags are also takes some time"*.

A user's strategy adoption habit was believed to have an influence on the number of resources that he or she would find during each visit. Specifically speaking, searchers would only find a few, whereas the adopters of other strategies, especially browsers, would find a lot. A participant provided a well-grounded explanation of such point of view:

- *"If I understand it correctly, users will use Douban search when they have a clear goal in mind… once they reach that goal, they don't need to find more… however, if a user does not like searching, it's probable that he likes to discover new things… because he doesn't know what exactly he is looking for, he will look at*

many resources out of curiosity".

Nevertheless, some participants did not think that strategy adoption was able to influence the number of resources the user would collect during each visit:

◆ "*Whether to collect a resource depends on many subjective factors…*"

When discussing the quantities of tags, contacts, or groups associated with different strategy adopters, the participants indicated additional characteristics of tag browsers, i.e. they would definitely attach importance to tagging and thus have many tags:

◆ "*This is a natural thing… if a user prefers to use tags to look for resources, he must know the importance of tagging the resources he has collected… his tagging will help other users find these resources*";

Although social monitors depended on their contacts or groups for resources, they would not necessarily have added many contacts or joined many groups:

◆ "*Even if I only observe one contact, as long as he often updates his collections, I will still have my needs satisfied*".

4.3.4 Interpretations of track level analysis results

After devoting a lengthy and full discussion to the issues concerning the different information seeking strategies, the participants in the focus group moved on to Q 6 and Q 7 that invited them to

probe into the major findings obtained from the clickstream data analysis at the track level. These two key questions respectively drew out 16 and 15 comments. The participants basically agreed with the findings presented and interpreted them based on their own understandings.

In the first place, the participants took it for granted that the more the webpages accessed during a visit, the more the resource pages that would be accessed:

◆ "*This is for sure… we come to Douban for resources*".

But this strong positive relationship between track length and capacity was not without exception:

◆ "*This may not be true to new users… when I was still new to Douban, I would access hundreds of webpages each time… besides looking for resources, I also wanted to acquaint myself with other features provided by Douban, such as discovering interesting people… sometimes most of the pages I viewed turned out be other people's photos*".

According to other comments, it was less safe to say that the more the webpages accessed during a visit, the more the resources that would be collected. The participants brought forward some reasons for not collecting interesting resources, such as:

◆ "*You need to sign in your account before collecting resources… I seldom do this when using public computers*";
◆ "*I may revisit the resources I have already added to my library to look at new reviews*".

When it came to the amount of time spent on a visit, the participants attributed its weaker

relationship with the numbers of resources found and collected during that visit to various causes. For examples:

◆ "*I spent the most time in writing reviews and reading others' reviews*";

◆ "*I've bookmarked Douban and it is one of the websites that will open automatically when I start the Web browser… I usually leave it open until I shut down my computer so that I can take a look at it from time to time*";

◆ "*Sometimes it only takes a couple of minutes to find a good album, but I will keep trial listen to the songs all day long*";

◆ "*How did you measure the time… did you deduct the time when users were away for dinner*".

Overall, the long-duration visits could be due to irrelevant yet time-consuming activities or users' inactivity in the website.

The participants identified three primary factors with an impact on the formation of resource finding patterns. First of all, if a user had a very specific information need, it was likely he or she would involve in sporadic finding. For examples:

◆ "*I think a user will only view one resource when he performs a search for a known resource… if I search with 'Harry Potter and the Deathly Hallows', this movie will appear on the top of the search results… I don't need to look at the rest*";

◆ "*Not necessarily searching… you do not have to remember the whole title… when you view the resources under the tag 'harry potter', you can also recognize the one you want*";

Secondly, the user's lower familiarity with a source was more likely to result in continuous finding. For example:

- *"If you are used to keep an eye on someone, each time you may only focus on a couple of new resources just added to his collection… but when you discover a new user, you will need to check out more resources in order to figure out his interests"*.

Last but not least, the continuous pattern was also associated with the user's higher evaluation of a source. For examples:

- *"Nobody will view a lot of resources from a poor source"*;
- *"How can you judge the usefulness of the resources recommended by Douban without looking them through"*.

Another way to understand these comments is that continuous finding actually reflected the processes of getting familiar with or evaluating a source aggregating multiple resources.

As for the formation of resource collecting patterns, the participants principally ascribed whether a user would collect a resource thoughtfully or thoughtlessly to his or her personal habit. Here are two contrasting examples:

- *"I only collect resources with high ratings, which means I think much of what others say about them"*;
- *"I use Douban mainly to keep track of the movies that I've watched… so when I collect a movie, I don't need to read its information"*;

The thoughtless collecting, moreover, could also be explained with some external influences, such as lack of attention to the resource and lack of time to view its details. For examples:

- "*Many resources are not frequently collected or widely reviewed… there is not much to read on their pages*";

- "*I will add several potentially interesting resources to my library every time and take a closer look at each one when I have spare time*".

5.0 DISCUSSION AND CONCLUSION

This dissertation study aimed at understanding users' information seeking behavior in the context of social tagging systems. It has generated both quantitative and qualitative results as reported in Chapter 4. In Section 5.1, the major results are discussed in terms of the three research questions of the study. The research questions were as follows:

1. What are the general information seeking strategies adopted by users in social tagging systems and how effective are they in helping users find information resources of interest?

2. For each information seeking strategy identified, is it possible to generalize the characteristics of the users who prefer to adopt it? If yes, what are these characteristics?

3. What are the specific traits of users' information seeking paths in social tagging systems and what are the factors contributing to the formation of their information seeking paths?

Section 5.2 presents a model created based on the empirical results. It is subsequently followed by the discussion of implications and suggestions for future research.

5.1 DISCUSSION OF MAJOR RESULTS

RQ 1: What are the general information seeking strategies adopted by users in social tagging systems and how effective are they in helping users find information resources of interest?

The clickstream data analysis at the movement level identified eight general information seeking strategies: encountering on home, browsing by resource, browsing by tag, browsing by user, browsing by group, searching, monitoring by user, and monitoring by group. They have their roots in the strategical ISB theories (Bates, 2002; Wilson, 1997), but develop in the context of social tagging systems. As a matter of fact, the universal tagging elements only include resources, tags, and users (Smith, 2008). However, this study also took into account two functional design elements, homepage and groups, that have become more and more standard in the construction of social tagging systems during the past half decade.

Firstly, the homepage designs of the systems now thinks less of the navigational purposes and instead pays more attention to content aggregation for users' convenience. Secondly, the designs of social interaction to be supported in the systems, in addition to social networking service, also consider groups which allow users to share information on common interests. Such changes have taken place or are taking place in most systems, and they show profound influences on users' information seeking behavior. As a whole, the ways users look for information in social tagging systems are greatly diversified in virtue of the connectivity among home, resources, tags, users, and groups, as illustrated in Figure 4.

Experimental research of encountering is difficult to design because it's hard to anticipate

who will acquire information in this way, where they will acquire information, or what information they will acquire (Erdelez, 2004). Such uncertainties are less obvious in the setting of social tagging systems. Being more social-oriented, they deliberately push information resources to users on their homepages, the common place to everyone. However these resources are usually limited and will be updated frequently. If one can find a resource of interest on the homepage, therefore, it is completely opportunistic.

Although resources can be encountered elsewhere, e.g. we may run across a resource when reading a group discussion topic making reference to it, they are actually ignorable compared to those encountered on the homepages. As uncovered in the clickstream data analysis, encountering on home was quite popular among Douban users, accounting for 17% of all the resource finding occurrences, which was the third highest. The great popularity of this strategy will probably be seen in other social tagging systems, considering that the visits to any websites usually start from the homepages. Consciously or unconsciously, users will notice the potentially interesting resources appearing there.

Meanwhile, the encountering strategy was quite effective in helping users find their needed resources, with the second highest find-to-collect rate (27.19%). The focus group participants attributed its high effectiveness mainly to Douban's success. That is to say, such result might be specific to this social tagging system only. It is true that Douban has been devoting a lot of efforts to resource recommendation. It carefully selects hundreds of recent, popular, and quality resources, and presents them to the users in a systematic manner. So in a system that does not have a comparable abundance of resources and/or lacks organization of resources on its homepage, the

effectiveness of this strategy might not be that high.

Browsing in social tagging systems sometimes is not clearly distinguishable from encountering because browsers also feel that they acquire information effort free. For example, on Douban's resource pages, the co-collected resources, if there are any, are just one click away. Notwithstanding, browsing differs from encountering for involving a proxy (McKenzie, 2003), being it a resource, a tag, a user, or a group. If a user is about to view the resources associated with a proxy, he or she is aware that they should be related to the proxy in some way. Although the user does not have a particular goal in mind, the subject or interest of the proxy represents his or her information need to certain extent. On the contrary, encountering is viewing resources not associated with any proxy.

Among the eight information seeking strategies identified, browsing by resource helped the users find 33% of the resources they ever found, which made it the most popular strategy. It is the most straightforward approach to acquiring related resources and takes two forms in Douban, browsing co-collected resources and browsing user-compiled lists of similar resources. Nevertheless, browsing related resources is not a ubiquitous strategy. It is mostly supported in social library systems, and not all of them support both forms, e.g. Discogs does not support the former. In spite of its popularity in Douban, this strategy had a find-to-collect rate (18.69%) lower than the average rate (22.38%) of all the strategies, suggesting unsatisfactory effectiveness. Especially the former form, according to the focus group, would often lead users to resources already viewed or collected before. This problem can also be found in other systems allowing users to browse based on relevance, such as LibraryThing.

In contrast, browsing by tag was the most effective strategy among the eight, though only demonstrating moderate popularity. As mentioned in Section 2.3, users tag resources in order to find them again later and help others discover them (Trant, 2009). Following tags to acquire resources, so to speak, is the most intrinsic information seeking strategy in social tagging systems. But the clickstream data analysis showed that it was only the fourth most frequently adopted strategy. Now one cannot say whether the strategy is less popular in other systems too, because a focus group participant thought that Douban users might be reluctant to use the tag cloud due to its low usability, which was a special problem in this system. Tags have attracted many doubts about their findability since they started to gain prevalence on the Web (Morvill, 2006). However it was found that the find-to-collect rate of browsing by tag reached as high as 33.31%, meaning that in every three resources found via tags, one of them would be collected. In that tags are semantic expressions, further investigation is needed to reveal if tags in other languages also have high findability.

Compared to the dominant role of Web search engines in general information seeking, the internal search engines provided by social tagging systems are affecting their users much less significantly in resource finding. In the case of Douban, the searching strategy failed to win overwhelming adoption, ranking the second in terms of popularity, and moreover, its find-to-click rate (23.42%) was only the third highest, implying merely acceptable effectiveness. It was mentioned several times in the focus group discussion that this strategy was mainly appropriate for tasks with specific goals. The disadvantages of Douban's search engine are in fact very common in other social tagging systems, such as Flickr, IMDb, and so forth. It is not surprising that the

recognizable search keywords are limited and the search results lack ranking. Interestingly, these are just trivial problems when the search engines are used for known item search.

The remaining four strategies, i.e. browsing by user, browsing by group, monitoring by user, and monitoring by group, are all characteristic of information seeking by social proxy, which is when users look for resources through an intermediary who is a particular person or a cluster of similar persons. Users and groups, as proxies, are not very different from each other. Both of them are describable with major interests, and the subjects of their collected resources should be able to reflect such interests. The browsing and monitoring strategies however work in different manners, with the former associated with newly discovered or unfamiliar users or groups and the latter those that people have established long-term relationships with. Before one starts to monitor a user or a group, he or she usually needs to do browsing first so as to determine whether it is a useful information source.

Based on the results of the clickstream data analysis, these four social-oriented strategies were neither popular nor effective. They together only explained a little more than 10% of all the occurrences of resource finding and their find-to-click rates (Mean = 12.70%) were far below the average level. When discussed in the focus group, they were either not considered as formal strategies or thought to be applicable only to users who had a passion for social activities. Social tagging systems, after all, are not social networking services, e.g. Facebook[31] and LinkedIn[32], which connect people who are real-world acquaintances and enable them to meet new friends through the

---

31 http://www.facebook.com/
32 http://www.linkedin.com/

old ones. The first and foremost goal here is finding resources of interest, and the finding of users or groups of interest is the byproduct. In addition, browsing or monitoring a user/group's collections is usually interwoven with browsing or monitoring that user/group's other information or updates. That is to say, people can be easily distracted from information seeking when adopting these strategies.

RQ 2: For each information seeking strategy identified, is it possible to generalize the characteristics of the users who prefer to adopt it? If yes, what are these characteristics?

The online survey investigated six types of information seekers in Douban whose favorite strategy respectively were encountering, browsing by resource, browsing by tag, searching, browsing by user/group, and monitoring by user/group. Being one of these types did not mean the exclusion of adopting other strategies. Most users categorized themselves as searchers (34.82%), followed by encounterers (19.64%), resource browsers (15.18%), tag browsers (13.39%), social monitors (9.82%), and social browsers (7.14%). It was found that they did not show significant differences in demographics, Web experience, search preference, system usage, user profile, or satisfaction level. However there were three exceptions.

First of all, encounterers tended to be new users with a visiting history in Douban shorter than 6 months ($p = .027$; $z = -2.9$, $z = 2.9$). The focus group participants thought that new users had two major disadvantages compared to the experienced ones: they were not familiar with all the available ways to look for resources and they had not built up their user profiles. For one thing, the encountering strategy, obviously, is the one with the lowest threshold, demanding for zero prior

familiarity or information seeking skill. Before getting used to other strategies that may better address their information needs, users have to depend heavily on encountering. For another, when still new to the system, users are in the beginning phase of enriching their libraries and establishing social connections. They may have insufficient knowledge of the proxy roles of the four basic elements in the system – resources, tags, users, and groups. Correspondingly, they will be less likely to adopt the strategies involving these elements.

As mentioned above, encountering works well in Douban because of its well-constructed homepages. In other social tagging systems which fail to give equal attention to their homepages, there may be less encounterers, but it is still possible that these encounterers are mostly new users. As a matter of fact, the most important feature of new users is that they do not have enough awareness or ability to avoid encountering resources on the homepages where they feel most confident and confortable. Even if they find that this strategy lacks effectiveness, they cannot, like experienced users, easily switch to other strategies. This is a problem that should be noticed by such systems as LibraryThing where encounterable resources are limited to the 10 most recently added, reviewed, and/or rated ones and Connotea where millions of recent or popular resources are plainly displayed without selecting. Their new users may have very poor resource finding experience.

Secondly, resource browsers and searchers were the two groups of users showing distinct characteristics in resource viewing frequency during each visit. The former tended to view more resources, i.e. more than 5 resources ($p = .010$; $z = 2.4$, $z = -2.4$), while the latter less, i.e. no more than 5 ($p = .010$; $z = -3.0$, $z = 3.0$). In certain sense, such results are expected by social

tagging systems. On the one hand, their resource pages are designed to display as many related resources as possible. Douban provides 10 co-collected resources and 5 user-compiled lists of similar resources for each resource. Some other systems will provide even more. For instance, LibraryThing includes system combined, member, and special recommendations on every resource page, and for popular resources, these related ones can add up to 100. On the other hand, social tagging systems usually restrict the functionalities of their internal search engines so that they are mostly used for known item search. This means that there will not be many resources returned on the search result pages.

However, the focus group participants thought that cognitive styles of resource browsers and searchers also contributed to these characteristics. The literature review has mentioned two different cognitive styles: field-dependence and field-independence (Kim & Allen, 2002). Resource browsers tended to be field-dependent. They are very interpersonal and have a well-developed ability to read social cues such as "people like this also like", and they are opener to follow these cues. But the potential risk is that, due to a lack of separation between the self and the field, they may easily get lost in the hyperlinks or feel great pressure when dealing with the considerable amount of information. By contrast, searchers tended to be field-independent. They are impersonal and task-oriented. When confronted with the search results, they will be reflective and cautious in order to avoid excessive input. Namely, they will go straightly to the most relevant result item(s) judged based on their own rationales, ignoring the possibility that a less relevant item may also be helpful.

Thirdly, tag browsers tended to have a large number of tags, larger than 50 ($p = .047$; $z =$

2.6, z = -2.6). It was believed by the focus group participants that tag browsers, with a deeper understanding of social tagging, had an internal impetus to tag. The folksonomy is built upon the collaborative efforts of many users. If nobody tags, no folksonomy exists and the strategy of browsing by tag will not work at all. Tag browsers may feel greater responsibility for contributing to a useful folksonomy because they benefit more from it than other strategy adopters. What should be noticed is that the large number of tags can be translated into the variety of user interests or the inconsistency of tag usage. When one has collected resources from a wide range of topic domains, he or she certainly need a lot of tags to describe them. But it's also possible that a topic is described with multiple tags which are the different forms of the same term.

RQ 3: What are the specific traits of users' information seeking paths in social tagging systems and what are the factors contributing to the formation of their information seeking paths?

This research question was explored in the clickstream data analysis at the track level. A track refers to a navigation path including all the pages accessed and all the actions taken during a visit. The 37,844 tracks extracted were analyzed in terms of four attributes which were length, duration, capacity, and achievement. Focused tracks, whose capacity was one or above, represented users' information seeking paths. It was found, based on the 21,583 focused tracks, that length had strong relationships with both capacity and achievement, but duration was not strongly related to either of them. 158 focused tracks, whose length was two or above, were selected for visualizing on the pathway graph proposed by this study. The researcher detected, from the visualizations generated,

two resource viewing patterns – continuous and sporadic, and two resource collecting patterns – lagged and instant. These results were all discussed in the focus group.

Among the four attributes, as seen in Table 4, the strongest relationship was that between length and capacity (r = .889). It could be inferred that information seeking activities take up a similar proportion in all the visits. And according to the focus group participants, this proportion should be very high since the major purpose for visiting Douban was resource finding, which was echoed by the online survey results of Q 5. In other social tagging systems where less irrelevant activities are supported, e.g. photo sharing, blogging, etc., this proportion should be even higher and the relationship between the two attributes should be even stronger. The relationship between length and achievement, though still strong, was much weaker (r = .619). The participants' explanations about this were that many resources viewed would not be collected even if they were interesting, either because the users did not sign in or the resources had already been collected.

In contrast, there was no strong relationship between duration and capacity or achievement (r = .530; r = .346). Firstly, it is not possible for a user to spend time evenly on every webpage accessed during a visit. More time will be spent on the pages containing more content, such as review pages with long texts. Next, different users access the webpages at different paces. In the online survey, some respondents reported that they could view 16 to 30 pages within 1 minute but some others needed 30 minutes. The former obviously were quicker viewers, whereas the latter thorough viewers. Besides, it was mentioned by the focus group participants that during a visit the user might just stay inactive in the website for some time. The website server however would not stop counting the time as long as they user did not leave the website. As a whole, how much time

users spend on their visits to a social tagging system cannot indicate the degree of their involvement in information seeking activities.

The other series of results concerning users' information seeking paths were the specific patterns found on them. In social tagging systems, resources are always accessed from a certain information source, being it the homepage, the search result page, or the resource, tag, user, and group pages. The continuous and sporadic resource viewing patterns respectively refer to the phenomena that the user views multiple resources from a source and that he or she views only one resource from that source. There were various user factors influencing the numbers of resources viewed at a specific source, as discussed in the focus group, including users' information needs, and their familiarity with and evaluation of the source. And in particular, the behavior of examining multiple resources from a source can be deemed a process of getting familiar with or evaluating the source.

After a resource is viewed, the most direct way to tell whether it is of interest to the user is to see whether it has been collected. The problem of this way consists in that the user might collect the resource instantly, i.e. collecting it without spending enough time reading its details. Because of the easiness of making changes to one's collections enabled by the social tagging systems, the user can delete the resource conveniently if finding out that it is not what he or she wants at a later time. Considering that the collecting action might be a fake signal of usefulness, the effectiveness of the information seeking strategies calculated above could be not accurate. But it is difficult to take into account the deleting of unwanted resources since the researcher cannot predict when this will happen. Of course, there should be only a tiny minority of resource collecting occurrences involved

such situation. The focus group participants also pointed out other reasons for instant collecting, such as the lack of content on the resource pages and collecting a previously known resource.

5.2 CONCLUSION

Based on the major findings of this dissertation study, the researcher creates a model of social tagging system users' information seeking behavior containing two layers: strategic and tactic (Figure 23).
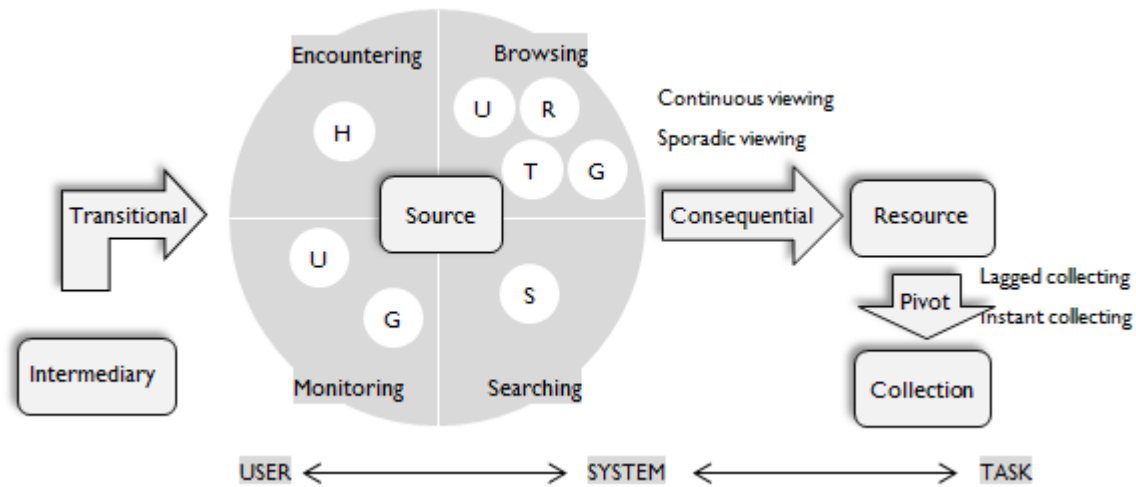
Figure 23. Model of social tagging system users' information seeking behavior

The strategic layer of the model is built upon the findings especially related to the first research question. It describes the high-level planning of information seeking activities by individual users in order to reach an information source, and shows how users select, apply, and monitor a strategy in the dynamic and complex environments of social tagging systems. As represented by the

circle area with darker shade in Figure 23, the strategic layer derives from Bates' (2002) model of information seeking and searching that comprises four different strategies: encountering, browsing, searching, and monitoring. The two strategies on the top tend to be more dependent upon the cues found in the environment, whereas the two on the bottom more goal driven. At the same time, while the two on the left just need the users to be available to absorb information, the two on the right require them to consciously act to seek information.

This strategic layer, in addition, subdivides the above strategies into eight more specific ones according to their interplay with the basic elements of social tagging systems: encountering on home (H), browsing by resource (R), browsing by tag (T), browsing by user (U), browsing by group (G), searching with search engine (S), monitoring by user (U), and monitoring by group (G). Because of these elements, resource finding in social tagging systems can be described as information seeking by proxy (McKenzie, 2003). By proxy refers to the occasions when users make direct contact with an information source. Except for searching, the elements in other sectors are not eternal. Influenced by the future development of social tagging systems, some elements may wither or even disappear, while some may become more prevalent and new ones will appear.

The tactic layer of the model originates from the findings generated for the third research question. It depicts the process in which a set of discrete intellectual choices made though behavioral actions during an information seeking session. The actions are conceptually uninteresting individually, but if taken in context, they assume different responsibilities. Transitional actions lead users to the information source where they can access the information resources. Although playing a less important role in an information seeking process, they inform us how users reach an

133

information source, thus indicating the strategies adopted. Consequential actions happen between the information source and the resource(s). They signal the potential usefulness of the resource(s), and can take place at the same source continuously or sporadically. In contrast, pivot actions signal the actual usefulness of the resource(s). The determination of the usefulness can happen after or without a time lag. When nothing is found useful, pivot actions will be absent from the whole process.

With the empirical findings of this dissertation study, it can be concluded that users behave very differently in social tagging systems from they do on the Web in general. According to a recent survey about people's everyday Web usage (TNS Global, 2008), 81% of the Web users rely on search engines to find information. However, the percentage of searchers in social tagging systems is less than 35%, as found in the online survey of this study. Users resort to various ways to find resources in social tagging systems, and these ways help them satisfy their information needs in varying degrees. Meanwhile, social tagging systems encourage users to explore and discover to a greater extent. Directed by their interests, users proceed with their navigation in the systems on less predictable paths. These paths take shape as the result of the collective effects of user, task, and system factors.

5.3 IMPLICATIONS

This dissertation study had three purposes. The first was concerned with the major information

seeking strategies adopted by users in social tagging systems. It was found that the strategies identified demonstrated highly different popularity and effectiveness. Especially, the most frequently adopted strategy, browsing by resource, was not very effective in helping users find their needed resources; however, the strategy with the highest find-to-collect rate, browsing by tag, only attracted moderate attention from the users. The second purpose of this study was to associate users' characteristics with their favorite strategies. The results indicate that individual differences among the users preferring different strategies were significant in certain aspects, including their familiarity with the system, resource finding habit, and involvement in the tagging activity. Finally, this study was also interested in the specific traits of users' information seeking paths in social tagging systems. As revealed by the analysis of the tracks representing the paths, longer visits to the system, during which more pages were accessed, tended to result in more occurrences of resource viewing and collecting. The model generated from these major findings provides useful implications for the design of user-oriented information seeking interfaces in general social tagging systems.

Social tagging systems are more diverse and dynamic information environments than the traditional Web, and people have been tailoring their behavior so as to keep optimum efficiency in information seeking. While search engines still serve an essential role, other approaches to resource finding either are gaining prevalence or have become very prevalent. For the sake of accommodating users' behavioral changes, all the basic architectural elements, including resource, tag, user, group, and home, etc., should be integrated in the development of social tagging systems.

In particular, the exploitation of tags needs to be promoted and facilitated, due to their remarkable usefulness in directing users to the resources of interest. It is not wise for such systems

as Discogs and IMDb to ignore the importance of offering users convenient access to tag clouds. Also deserving thoughtful consideration is the construction of system homepages, because encountering on home is an effortless yet satisfactory strategy which can be adopted by any users. The common practice is pushing the recently released, widely discussed, and highly rated information resources on the homepages. But the abundance of resources must be appropriate, or users will be defeated by information overload.

At present, a major functionality missing from most social tagging systems is the customizable control over the use of the systems by the users themselves. Take Douban's resource pages for example. With no exception, any resource will be presented together with its most frequently attached tags, its individual and group collectors, its co-collected resources, and so on. It is true that the variety of information seeking methods is ensured. But what if a user just wants one of the components, e.g. the tags for browsing by tag? This means that the other components are all noise and they could distract the user's attention from the component that he or she is really interested in.

One possible solution to such problem is dividing various components into independent modules and giving users the permission to deactivate and reactivate the modules according to their specific needs. That is to say, system designs will be more user-friendly when taking into account the individual differences. If a certain module fails to support users' information seeking preferences, it can be deactivated. Indeed, their preferences will not keep unvaried forever. For instances, new users' dependence on the homepage for encountering resources will be alleviated as they spend more time in a system and get more familiar with it, and the "people who like this also like"

component will become less helpful in recommending new resources as users' libraries grow larger. Consequently, the management of system modules should allow adequate flexibility to satisfy users' changing demands.

An additional lesson we learn from this study is about the range of services provided by social tagging systems. Despite that Douban focuses on enabling users to discover, collect, and tag books, movies, and music, it also involves other services for socializing purposes, such as publishing blogs and sharing photos. The latter services seem to add value to the system, but they are in fact weakening its image as a social library system. As seen in Figure 16, when track lengths increase, the distributions of capacity and achievement values become broader, however with most tracks aggregating at the low end. In other words, although the activities performed by users increase, their resource finding and collecting activities still stay at a low level. They may be mainly occupied in social activities during their visits to the system. Obviously, the importance of fundamental services and value-added services are reversed. Social tagging systems, hence, should restrict their expansion to social networking services to a moderate degree and devote to enhancing their findability as information seeking environments.

5.4 SUGGESTIONS FOR FUTURE RESEARCH

In spite of the extensive attention given to researching Web information seeking, this dissertation is one of the first to investigate users' information seeking behavior in the context of social tagging

systems. Given that these newly born systems are gaining prevalence on the Web and becoming important information repositories for users with different purposes, future research is suggested to further the investigation in this specific context.

A more generalizable study on this topic needs to occur. This study chose a language specific social library system as the research setting. Although Douban serves a super large number of users, the absolute majority of them belong to the Chinese-speaking world in which their information seeking behavior may be affected by the culture, society, and other factors. Besides, the social library system is one of the categories of social tagging systems, and the resources contained are limited to books, movies, and music albums. But the type of the resources may have an impact on how they will be found. In order to increase the generalizability of the results, research needs to be conducted with social tagging systems in other languages and/or with other types of resources.

It is also important for future research to address the problem of data collection appearing in this study. The advantage of using the transaction log file from the website server is that the clickstream data included reveals users' realistic behavior. However the lack of users' background information in the transaction log requires the researcher to collect it separately. Since there is no way to guarantee that users sampled in the two processes largely overlap, the connections found between behavior and background may be not persuasive enough. A possible solution to such problem is collecting the two kinds of data at the same time in a laboratory environment where strong internal validity is often gained at the expense of external validity (Krathwohl, 2004).

Aside from the research topic, the clickstream data analysis method can be pursued in the future to confirm its usefulness and suggest improvements. The original clickstream data analysis

framework was proposed by Sen *et al.* (2006) yet never implemented in any empirical studies. This study makes necessary changes to it by taking into account the traits of the clickstream data from social tagging systems. The applicability of the new framework in different information seeking systems needs to be tested.

APPENDIX A. TAXONOMY OF FOOTPRINTS IN DOUBAN

Note: the abbreviations in the URLs are read as follows: [rid] = resource ID; [uid] = user ID; [t] = tag; [q] = search keyword; [c] = category; [gid] = group ID; [vid] = review ID; [did] = discussion topic ID; and [lid] = doulist ID.

Category	URL	Description	Abbr.	Code
Home	/	Douban home	H	4.0
	/book/ /movie/ or /movie/tv /music/	Classified homes for books, movies, TV series, and music albums respectively	Hr	
	/book/chart /movie/chart /music/chart	A list of popular new resources	Hn	4.1
	/book/top250 /movie/top250 /music/top250	A list of top rated resources selected by Douban	Hp	4.2

	/book/review/(best, latest)/ /movie/review/(best, latest)/ /music/review/(best, latest)/	A list of popular or recent user reviews	Hv	4.3
	/book/recommended /movie/recommended /music/recommended	A list of personalized recommendations of resources provided by Douban	He	4.4
	/book/browse /movie/browse /music/browse /book/tag/ /movie/tag/ /music/tag/ /movie/cluster/[c] /music/cluster/[c] /music/tags/[c]	Tag clouds	Ht	4.5
	/group/ /group/discover	Home for interest groups	Hg	4.6
	/group/category/[c]	Classifications of interest groups		
Resource	/subject/[rid]/	A particular resource	R	5.0
	/subject/[rid]/collections /subject/[rid]/doings /subject/[rid]/wishes	A list of users who have collected a particular resource	Ru	5.1

	/subject/[rid]/reviews	A list of user reviews of a particular resource	Rv	5.2
	/subject/[rid]/discussion	A list of discussion topics about a particular resource	Rd	5.3
	/subject/[rid]/doulists	A list of doulists containing a particular resource	Rl	5.4
	/subject/[rid]/group_collectors	A list of interest groups which have collected a particular resource	Rg	5.5
	/review/[vid]/	A particular user review of a particular resource	V	5.6
	/subject/discussion/[did]	A particular discussion topic about a particular resource	D	5.7
Tag	/book/tag/[t] /movie/tag/[t] /music/tag/[t]	A list of resources associated with a particular tag	T	6.0
User	/people/[uid]/	A particular user	U	3.0
	/book/list/[uid]/ /movie/list/[uid]/ /music/list/[uid]/ /people/[uid]/booktags/[t] /people/[uid]/movietags/[t] /people/[uid]/musictags/[t]	A particular user's collections of resources	Ur	3.1
	/people/[uid]/friend_list	A particular user's contacts	Uf	3.2

	/people/[uid]/contact_list /people/[uid]/rev_contacts			
	/people/[uid]/reviews	Resource reviews submitted by a particular user	Uv	3.3
	/people/[uid]/doulists	Doulists created by a particular user	Ul	3.4
	/people/[uid]/groups	Interest groups that a particular user has joined	Ug	3.5
	/people/[uid]/recs	Recommendations made by a particular user	Ue	3.6
	/people/[uid]/miniblogs	A particular user's recent activities	Un	3.7
	/doulist/[lid]/	A particular doulist compiled by a user	L	3.8
Group	/group/[gid]/	A particular interest group	G	1.0
	/group/[gid]/collection	A list of resources collected by a particular group	Gr	1.1
	/group/[gid]/members	A list of users who have joined a particular interest group	Gu	1.2
"Mine" (only visible to the current signed-in user)	/mine/	My Douban library (personal home)	M	2.0
	/book/mine /movie/mine /music/mine	A list of my collected resources	Mr	2.1
	/contacts/list /contacts/listfriends	A list of my contacts	F	2.2

	/contacts/	A list of my contacts' recent activities	Fn	2.3
	/mine/discussions	A list of my discussion topics	Md	2.4
	/mine/collect_doulist	A list of my collected doulists	Ml	2.5
	/group/mine	A list of interest groups that I have joined	Mg	2.6
	/mine/recs	A list of my recommendations	Me	2.7
	/mine/miniblogs	A list of my recent activities	Mn	2.8
Search	/subject_search?search_text=[q] /book/search/[q] /movie/search/[q] /music/search/[q] /music/song_search?q=[q] /amazon_search?search_text=[q]	An action that submit a particular query to the search engine which returns a list of resources	S	7.0
Collect	/j/subject/[rid]/interest?interest=(collect, do, wish) /(collection, do, wish)/[uid]/update?add=[rid] /subject/[rid]/?interest=(collect, do, wish)	An action that adds a particular resource to a collection	C	8.0
Peripheral	-	Uncategorized	P	0.0

APPENDIX B. SQL QUERIES

Query 1

Select UID, mid (REQ, 10, 7) as RID, REQ, REF, TIME into resource_finding from cleaned_data where REQ like '/subject/*' and (len (REQ) < 18 or REQ like '*?from*' or REQ like '*?i*' or REQ like '*?rec*') and REQ not like '*interest*' and REQ not like '*discussion*';

Query 2

Select UID, mid (REQ, 12, 7) as RID, REQ, REF, TIME into resource_collecting from cleaned_data where REQ like '/j/subject/*' and REQ not like '*interest?';
Select UID, mid (REQ, 10, 7) as RID, REQ, REF, TIME into resource_collecting from cleaned_data where REQ like '/subject/*' and REQ like '*?interest*';
Select UID, right (REQ, 7) as RID, REQ, REF, TIME into resource_collecting from cleaned_data where (REQ like '/collection/*' or REQ like '/do/*' or REQ like '/wish/*') and REQ like '*update*';

Query 3

Select UID, count (UID) from resource_finding group by UID;

Query 4

Select UID, count (UID) from resource_collecting group by UID;

Query 5

Select RID, count (RID) from resource_finding group by RID;

Query 6

Select RID, count (RID) from resource_collecting group by RID;

Query 7

Select resource_finding.UID, resource_finding.REQ, resource_finding.REF, resource_finding.TIME from resource_finding inner join resource_collecting on resource_finding.UID = resource_collecting.UID and resource_finding.REQ = resource_collecting.REF and resource_finding.TIME < resource_collecting.TIME;

Query 8

Select cleaned_data.UID, cleaned_data.REQ, cleaned_data.REF, cleaned_data.TIME into regular_data from cleaned_data inner join registered_uids on cleaned_data.UID = registered_uids.UID;

Query 9

Select TID, count (TID) from regular_data group by TID;

Query 10

Select TID, count (TID) from regular_data where REQ like '/subject/*' and (len (REQ) < 18 or REQ like '*?from*' or REQ like '*?i*' or REQ like '*?rec*') and REQ not like '*interest*' and REQ not like '*discussion*' group by TID;

Query 11

Select TID, count (TID) from regular_data where REQ like '/j/subject/*' and REQ not like '*interest?' or REQ like '/subject/*' and REQ like '*?interest*' or (REQ like '/collection/*' or REQ like '/do/*' or REQ like '/wish/*') and REQ like '*update*' group by TID;

APPENDIX C. VBA MACROS

Macro *identify_tracks*

Public Sub identify_tracks()

```
Dim h, n
h = 2
n = 1

Do Until Sheet1.Cells(h, 1) = ""
   If Sheet1.Cells(h, 1) <> Sheet1.Cells(h - 1, 1) Then
   Sheet1.Cells(h, 5) = 1
   n = 1
   End If
   If Sheet1.Cells(h, 1) = Sheet1.Cells(h - 1, 1) And Sheet1.Cells(h - 1, 4) > DateAdd("s", -1800,
CDate(Sheet1.Cells(h, 4))) Then
   Sheet1.Cells(h, 5) = n
   End If
   If Sheet1.Cells(h, 1) = Sheet1.Cells(h - 1, 1) And Sheet1.Cells(h - 1, 4) <= DateAdd("s", -1800,
CDate(Sheet1.Cells(h, 4))) Then
   Sheet1.Cells(h, 5) = n + 1
   n = n + 1
   End If
   h = h + 1
Loop

End Sub
```

Macro *restore_first*

Public Sub restore_first()

Dim h
h = 2

Do Until Sheet1.Cells(h, 1) = ""
    If Sheet1.Cells(h, 1) <> Sheet1.Cells(h - 1, 1) Or Sheet1.Cells(h, 1) = Sheet1.Cells(h - 1, 1) And
Sheet1.Cells(h, 5) <> Sheet1.Cells(h - 1, 5) Then
        Sheet1.Cells(h, 6) = Sheet1.Cells(h, 1)
        Sheet1.Cells(h, 7) = Sheet1.Cells(h, 3)
        Sheet1.Cells(h, 8) = "-"
        Sheet1.Cells(h, 9) = DateAdd("s", -2, CDate(Sheet1.Cells(h, 4)))
        Sheet1.Cells(h, 10) = Sheet1.Cells(h, 5)
    End If
    h = h + 1
Loop

End Sub

Macro *insert_first*

Public Sub insert_first()

Dim h1, h2
h1 = 2
h2 = 2

```
Do Until Sheet1.Cells(h1, 1) = ""
    If Sheet1.Cells(h1, 6) <> "" Then
        Sheet2.Cells(h2, 1) = Sheet1.Cells(h1, 6)
        Sheet2.Cells(h2, 2) = Sheet1.Cells(h1, 7)
        Sheet2.Cells(h2, 3) = Sheet1.Cells(h1, 8)
        Sheet2.Cells(h2, 4) = Sheet1.Cells(h1, 9)
        Sheet2.Cells(h2, 5) = Sheet1.Cells(h1, 10)
        h2 = h2 + 1
        Sheet2.Cells(h2, 1) = Sheet1.Cells(h1, 1)
        Sheet2.Cells(h2, 2) = Sheet1.Cells(h1, 2)
        Sheet2.Cells(h2, 3) = Sheet1.Cells(h1, 3)
        Sheet2.Cells(h2, 4) = Sheet1.Cells(h1, 4)
        Sheet2.Cells(h2, 5) = Sheet1.Cells(h1, 5)
        h2 = h2 + 1
    Else
        Sheet2.Cells(h2, 1) = Sheet1.Cells(h1, 1)
        Sheet2.Cells(h2, 2) = Sheet1.Cells(h1, 2)
        Sheet2.Cells(h2, 3) = Sheet1.Cells(h1, 3)
        Sheet2.Cells(h2, 4) = Sheet1.Cells(h1, 4)
        Sheet2.Cells(h2, 5) = Sheet1.Cells(h1, 5)
        h2 = h2 + 1
    End If
        h1 = h1 + 1
Loop

End Sub
```

Macro *restore_interrupted*

Public Sub restore_interrupted()

```
Dim h1, h2, h3, i
h1 = 2
h2 = 1
i = 0


Do Until Sheet2.Cells(h1, 1) = ""
    If Sheet2.Cells(h1, 1) <> Sheet2.Cells(h1 - 1, 1) Or Sheet2.Cells(h1, 1) = Sheet2.Cells(h1 - 1, 1) And
Sheet2.Cells(h1, 5) <> Sheet2.Cells(h1 - 1, 5) Then
        h3 = h1
    End If
    If Sheet2.Cells(h1, 3) <> Sheet2.Cells(h1 - 1, 2) Then
      h2 = h1 - 1
      Do Until i = 1 Or h2 < h3
        If Sheet2.Cells(h1, 3) = Sheet2.Cells(h2, 2) Then
          i = 1
        End If
        h2 = h2 - 1
        If h2 = 0 Then
         h2 = 1
        End If
      Loop
    End If
    If i = 0 And Sheet2.Cells(h1, 3) <> Sheet2.Cells(h1 - 1, 2) And Sheet2.Cells(h1, 1) = Sheet2.Cells(h1 - 1,
1) And Sheet2.Cells(h1, 5) = Sheet2.Cells(h1 - 1, 5) Then
        Sheet2.Cells(h1, 6) = Sheet2.Cells(h1, 1)
        Sheet2.Cells(h1, 7) = Sheet2.Cells(h1, 3)
        Sheet2.Cells(h1, 8) = "-"
        Sheet2.Cells(h1, 9) = DateAdd("s", -2, CDate(Sheet2.Cells(h1, 4)))
        Sheet2.Cells(h1, 10) = Sheet2.Cells(h1, 5)
    End If
    i = 0
    h1 = h1 + 1
Loop

End Sub
```

Macro *insert_interrupted*

```
Public Sub insert_interrupted()

Dim h1, h2
h1 = 2
h2 = 2

Do Until Sheet2.Cells(h1, 1) = ""
    If Sheet2.Cells(h1, 6) <> "" Then
        Sheet3.Cells(h2, 1) = Sheet2.Cells(h1, 6)
        Sheet3.Cells(h2, 2) = Sheet2.Cells(h1, 7)
        Sheet3.Cells(h2, 3) = Sheet2.Cells(h1, 8)
        Sheet3.Cells(h2, 4) = Sheet2.Cells(h1, 9)
        Sheet3.Cells(h2, 5) = Sheet2.Cells(h1, 10)
        h2 = h2 + 1
        Sheet3.Cells(h2, 1) = Sheet2.Cells(h1, 1)
        Sheet3.Cells(h2, 2) = Sheet2.Cells(h1, 2)
        Sheet3.Cells(h2, 3) = Sheet2.Cells(h1, 3)
        Sheet3.Cells(h2, 4) = Sheet2.Cells(h1, 4)
        Sheet3.Cells(h2, 5) = Sheet2.Cells(h1, 5)
        h2 = h2 + 1
    Else
        Sheet3.Cells(h2, 1) = Sheet2.Cells(h1, 1)
        Sheet3.Cells(h2, 2) = Sheet2.Cells(h1, 2)
        Sheet3.Cells(h2, 3) = Sheet2.Cells(h1, 3)
        Sheet3.Cells(h2, 4) = Sheet2.Cells(h1, 4)
        Sheet3.Cells(h2, 5) = Sheet2.Cells(h1, 5)
        h2 = h2 + 1
    End If
            h1 = h1 + 1
  Loop

End Sub
```

Macro *track_duration*

Public Sub track_duration()

Dim h
h = 2

```
Do Until Sheet5.Cells(h, 1) = ""
    If Sheet5.Cells(h, 3) = "exit" Then
        Sheet5.Cells(h, 5) = DateDiff("s", CDate(Sheet5.Cells(h - 1, 1)), CDate(Sheet5.Cells(h, 1)))
    End If
        h = h + 1
Loop
```

End Sub

Macro *repliate_multitask*

Public Sub repliate_multitask ()

```
Dim h1, h2, h3, i
h1 = 2
h2 = 1
i = 0

Do Until Sheet3.Cells(h1, 1) = ""
     If Sheet3.Cells(h1, 1) <> Sheet3.Cells(h1 - 1, 1) Then
         h3 = h1
     End If
     If Sheet3.Cells(h1, 3) <> Sheet3.Cells(h1 - 1, 2) Then
       h2 = h1 - 1
       Do Until i = 1 Or h2 < h3
         If Sheet3.Cells(h1, 3) = Sheet3.Cells(h2, 2) Then
            Sheet3.Cells(h1, 5) = Sheet3.Cells(h2, 1)
            Sheet3.Cells(h1, 6) = Sheet3.Cells(h2, 2)
            Sheet3.Cells(h1, 7) = Sheet3.Cells(h2, 3)
            Sheet3.Cells(h1, 8) = Sheet3.Cells(h2, 4)
             i = 1
         End If
         h2 = h2 - 1
         If h2 = 0 Then
           h2 = 1
         End If
       Loop
     End If
     i = 0
     h1 = h1 + 1
Loop
End Sub
```

Macro *insert_multitask*

Public Sub insert_multitask()

Dim h1, h2
h1 = 2
h2 = 2

```
Do Until Sheet3.Cells(h1, 1) = ""
    If Sheet3.Cells(h1, 5) <> "" Then
        Sheet4.Cells(h2, 1) = Sheet3.Cells(h1, 5)
        Sheet4.Cells(h2, 2) = Sheet3.Cells(h1, 6)
        Sheet4.Cells(h2, 3) = Sheet3.Cells(h1, 7)
        Sheet4.Cells(h2, 4) = Sheet3.Cells(h1, 8)
        Sheet4.Cells(h2, 5) = "1"
        h2 = h2 + 1
        Sheet4.Cells(h2, 1) = Sheet3.Cells(h1, 1)
        Sheet4.Cells(h2, 2) = Sheet3.Cells(h1, 2)
        Sheet4.Cells(h2, 3) = Sheet3.Cells(h1, 3)
        Sheet4.Cells(h2, 4) = Sheet3.Cells(h1, 4)
        h2 = h2 + 1
    Else
        Sheet4.Cells(h2, 1) = Sheet3.Cells(h1, 1)
        Sheet4.Cells(h2, 2) = Sheet3.Cells(h1, 2)
        Sheet4.Cells(h2, 3) = Sheet3.Cells(h1, 3)
        Sheet4.Cells(h2, 4) = Sheet3.Cells(h1, 4)
        h2 = h2 + 1
    End If
            h1 = h1 + 1
  Loop
```

End Sub

Macro *visualize_track*

Public Sub visualize_track()

```
Dim h
Dim h1
Dim h2
Dim db, de
Dim eb, ee
Dim i
h = 1
i = 0

Charts.Add
ActiveChart.Location Where:=xlLocationAsObject, Name:="Sheet6"
    ActiveChart.Axes(xlValue).MinimumScale = 0
    ActiveChart.Axes(xlValue).MaximumScale = 8
    ActiveChart.Axes(xlValue).MajorUnit = 1

With ActiveChart
  .HasLegend = True
  .ChartType = xlXYScatterLinesNoMarkers
  .HasTitle = True
  .ChartTitle.Text = CStr(Sheet6.Cells(h, 1))
End With

With ActiveChart.Axes(xlCategory)
  .HasTitle = True
  .AxisTitle.Text = "time"
  End With

With ActiveChart.Axes(xlValue)
  .HasTitle = True
  .AxisTitle.Text = "category"
End With

Do Until Sheet6.Cells(h, 1) = ""
    If Sheet6.Cells(h, 6) = 1 Then
        h1 = h
        h = h + 1
        Do Until Sheet6.Cells(h, 6) = 1 Or Sheet6.Cells(h, 1) = ""
```

```vba
        h = h + 1
      Loop
      h2 = h - 1
      i = i + 1
         db = "d" & h1
         de = "d" & h2
         eb = "e" & h1
         ee = "e" & h2
    ActiveChart.SeriesCollection.NewSeries
    ActiveChart.SeriesCollection(i).XValues = Sheet6.Range(db & ":" & de)
    ActiveChart.SeriesCollection(i).Values = Sheet6.Range(eb & ":" & ee)
    ActiveChart.SeriesCollection(i).Border.Color = RGB(70, 130, 180)
    ActiveChart.SeriesCollection(i).Select
    With Selection.Format.Line
         .Visible = msoTrue
         .Weight = 0.25
    End With
    Selection.Format.Line.EndArrowheadStyle = msoArrowheadStealth
    End If
Loop

End Sub
```

APPENDIX D. ONLINE SURVEY QUESTIONNAIRE FORM

Douban User Survey

As part of the research project studying users' information seeking behavior in the context of a typical social tagging system – Douban (http://www.douban.com/), this survey aims to collect relevant information about the users. All respondents should be registered Douban users who visit the website regularly and will be asked to complete a brief (10-15minutes) Web-based questionnaire with 21 close-ended multiple choice questions.

If you are willing to participate, you will need to describe (1) your profile as a Douban user, i.e. your usage of the system; (2) your information seeking experience in Douban, such as how you look for resources, and whether you are satisfied with what you find; and (3) your background information as a general Web user, including your demographics, Web expertise, and search preferences.

This is an anonymous survey, and your responses are totally confidential and will be exclusively used in this research project. Your participation is voluntary, and you may withdraw from the survey at any time. If you have any questions, please contact the researcher at doubanresearch@gmail.com.

--PAGE 1--

1. How long have you been visiting Douban?
- ☐ ○   Less than 3 months
- ☐ ○   3 months ~ less than 6 months
- ☐ ○   6 months ~ less than 1 year
- ☐ ○   1 year ~ less than 3 years
- ☐ ○   3 years or more

2. How often do you visit Douban?

☐ ○   More than once a day

☐ ○   Daily

☐ ○   Weekly

☐ ○   Monthly

☐ ○   Seldom

3. How many webpages do you usually view each time you visit Douban?

☐ ○   ≤ 5

☐ ○   6 ~ 15

☐ ○   16 ~ 30

☐ ○   31 ~ 50

☐ ○   > 50

4. How much time do you usually spend on Douban during each visit?

☐ ○   Less than 1 minute

☐ ○   1 minute ~ less than 10 minutes

☐ ○   10 minutes ~ less than 30 minutes

☐ ○   30 minutes ~ less than 2 hours

☐ ○   2 hours or more

5. Why do you visit Douban? (Please select all that apply)

☐ ☐   Discovering new books, movies, or music albums that I don't know

☐ ☐   Collecting books, movies, or music albums that I've heard elsewhere

☐ ☐   Social networking, i.e. meeting friends, participating in interest groups, etc.

☐ ☐   Using other services provided by Douban, e.g. blogs, photo sharing, e-mail, etc.

☐ ☐   No specific purpose

☐ ☐   Other (please specify) _____

6. How many resources (books + movies + music albums) have you collected in Douban?

☐ ○   ≤ 10

☐ ○   11 ~ 50

☐ ○   51 ~ 100

☐ ○   101 ~ 200

☐ ○   > 200

7. How many tags (book tags + movie tags + music tags) have you created in Douban?

☐ ○   ≤ 10
☐ ○   11 ~ 50
☐ ○   51 ~ 100
☐ ○   101 ~ 200
☐ ○   > 200

8. How many contacts (your friends + those you observe) do you have in Douban?

☐ ○   ≤ 10
☐ ○   11 ~ 50
☐ ○   51 ~ 100
☐ ○   101 ~ 200
☐ ○   > 200

9. How many interest groups have you joined in Douban?

☐ ○ ≤ 2
☐ ○ 3 ~ 5
☐ ○ 6 ~ 10
☐ ○ 11 ~ 20
☐ ○ > 20

--PAGE 2--

10. How many resources (books, movies, or music albums) do you usually find each time you visit Douban?

☐ ○   ≤ 2
☐ ○   3 ~ 5
☐ ○   6 ~ 15
☐ ○   16 ~ 30
☐ ○   > 30

11. How many resources (books, movies, or music albums) do you usually collect each time you visit Douban?

☐ ○   0
☐ ○   1 ~ 2

☐ ○ 3 ~ 5
☐ ○ 6 ~ 10
☐ ○ > 10

12. What method(s) do you use to look for resources in Douban? (Please select all that apply)

☐ ☐ Using the internal search engine
☐ ☐ Following the resources recommended on the homepages
☐ ☐ Scanning the tag clouds
☐ ☐ Viewing "people who like this also like" or doulists
☐ ☐ Exploring the resource collections of random users or groups that I come across
☐ ☐ Observing the updates in the resource collections of my contacts or affiliated groups
☐ ☐ Other (please specify) _____

13. Which method among the above do you use most frequently?

☐ ○ Using the internal search engine
☐ ○ Scanning the resources recommended on the homepages
☐ ○ Following the tags
☐ ○ Viewing "people who like this also like" or doulists
☐ ○ Exploring the resource collections of random users or groups that I come across
☐ ○ Observing the updates in the resource collections of my contacts or affiliated groups
☐ ○ Other

14. It is possible that a resource you find in Douban is not what you need. How often does this happen to you?

☐ ○ Never
☐ ○ Seldom
☐ ○ Occasionally
☐ ○ Frequently
☐ ○ Constantly

--PAGE 3--

15. What is your age?

☐ ○ ≤ 18
☐ ○ 19 ~ 22
☐ ○ 23 ~ 30

163

☐ ○   31 ~ 40

☐ ○   > 40

16. What is your gender?

☐ ○   Female

☐ ○   Male

17. What is your highest level of education?

☐ ○   High school

☐ ○   College degree

☐ ○   Bachelor's degree

☐ ○   Master's degree

☐ ○   Doctoral degree

☐ ○   Other

18. How long have you been using the Web to look for information?

☐ ○   Less than 6 months

☐ ○   6 months ~ less than 1 year

☐ ○   1 year ~ less than 2 years

☐ ○   2 years ~ less than 5 years

☐ ○   5 years or over

19. How often do you use the Web to look for information?

☐ ○   More than once a day

☐ ○   Daily

☐ ○   Weekly

☐ ○   Monthly

☐ ○   Seldom

20. What method(s) do you use to look for information on the Web? (Please select all that apply)

☐ ☐   Search engines

☐ ☐   Web directories

☐ ☐   Web portals

☐ ☐   Bookmarked websites

☐ ☐   Other (please specify) _____

21. Which method among the above do you use most frequently?

□ ○   Search engines

□ ○   Web directories

□ ○   Web portals

□ ○   Bookmarked websites

□ ○   Other

---END OF SURVEY---

* Are you interested in participating in a related focus group?

As a participant of the focus group, you will help with explaining a series of research results concerning Douban users' information seeking behavior based on your own experience and understanding. You will need to communicate, in an anonymous manner, with the researcher and 5 other focus group participants via instant messaging software, e.g. Google Talk. All the participants will be asked to express their opinions on 7 predefined questions and discuss with each other. The whole session is expected to last 1.5 to 2 hours.

If you are interested in participating in this focus group, please leave your Email address below. The researcher (*doubanresearch@gmail.com*) will contact you to set up the session. Thank you! Your Email address: _____

APPENDIX E. FOCUS GROUP QUESTIONING ROUTE

1. Tell us how long you have been a Douban user, what you visit Douban for, and whether you like the website.

2. Think back to the early days of your visit to Douban. Have you since then developed any specific habits as far as resource finding and collecting are concerned? If yes, what are they?

   The following are the major methods used by users to look for resources in Douban:
   a) Using the internal search engine
   b) Following the resources recommended on the homepages
   c) Scanning the tag clouds
   d) Viewing "people who like this also like" or doulists
   e) Exploring the resource collections of random users or groups that I come across
   f) Observing the updates in the resource collections of my contacts or affiliated groups

3. Think about the method(s) that you are familiar with. How frequently is it (are they) used by Douban users? Can you talk about the reasons for that the method(s) is (are) more or less frequently used than others?

4. After viewing the details about the resources found through a method, users may decide whether to collect them. So every method has a find-to-click rate. For example, if 100 resources are found using a particular method but 50 of them are collected finally, then this method's find-to-collect rate is 50%. Which methods do you think have higher rates and which ones have lower rates? And your reasons?

5. Do the users preferring the same method share similar characteristics, such as age,

gender, education level, web using history and frequency, search preference, Douban visiting history and frequency, the number of webpages accessed during each visit to Douban, the time spent on each visit, the number of resource viewed and collected during a visit, the numbers of associated resources, tags, contacts, and groups, and resource finding satisfaction level? Please support your comments.

6. As found in this study, during a visit to Douban, the number of webpages accessed was strongly and positively related to both the numbers of resources viewed and collected. But the amount of time spent on the visit failed to show a strong relationship with either of them. Can you explain both phenomena?

7. When a user reaches a source in Douban which may lead him or her to more than one resource, e.g. a user or a group's resource collections, are there any particular factors determining or influencing the numbers of resources the user will click through? Before collecting a resource, are there any particular factors determining or influencing how much time the user needs to decide that the resource is worth collecting?

BIBLIOGRAPHY

Astrom, F. (2007). Changes in the LIS research front: Time-sliced cocitation analyses of LIS journal articles, 1990-2004. Journal of the American Society for Information Science and Technology, 58 (7), 947-957.

Adamic, L. A. (2000). Zipf, power-law, pareto – a ranking tutorial. Retrieved October 29, 2010, from http://www.hpl.hp.com/research/idl/papers/ranking/ranking.html

Aspelmeier, J., & Pierce, T. (2009). SPSS: user friendly approach. Worth Publishers.

Baeza-Yates, R., & Castillo, C. (2001). Relating web structure and user search behavior. In Proceedings of the 10th World Wide Web Conference, 1-2.

Bates, M. J. (1989). The design of browsing and berrypicking techniques for the online search interface. Online Review, 13 (5), 407-424.

Bates, M. J. (2002). Toward an integrated model of information seeking and searching. Keynote of the 4th International Conference on Information Needs, Seeking and Use. Retrieved March 25, 2010, from http://www.gseis.ucla.edu/faculty/bates/articles/info_SeekSearch-i-030329.html

Bates, M. J. (2005). An introduction to metatheories, theories, and models. In K. E. Fisher, S. Erdelez, & L. McKechnie (Ed.), Theories of information behavior (pp. 1-24). Medford, NJ: Information Today, Inc.

Beitzel, S. M., Jensen, E. C., Chowdhury, A., Grossman, D., & Frieder, O. (2004). Hourly analysis of a very large topically categorized Web query log. In Proceedings of the 27th Annual International Conference on Research and development in Information Retrieval, 321–328.

Berstam, E. V., Herskovic, J. R., & Hersh, W. R. (2008). Query log analysis in Biomedicine. In B. J. Jansen, A. Spink, & I. Taksa (Ed.), Handbook of research on Web log analysis (pp. 359-377). IGI Global.

Blecic, D. D., Bangalore, N. S., Dorsch, J. L., Henderson, C. L., Koenig, M. H., & Weller, A. C. (1998). Using transaction log analysis to improve OPAC retrieval results. College & Research Libraries, 59 (1), 39-50.

Booth, D., & Jansen, B. J. (2008). A review of methodologies for analyzing websites. In B. J. Jansen, A. Spink, & I. Taksa (Ed.), Handbook of research on Web log analysis (pp. 143-164). IGI Global.

Bradley, P. (2007). How to use Web 2.0 in your library. London: Facet Publishing.

Bronstein, J. (2007). The role of the research phase in information seeking behavior of Jewish studies scholars: A modification of Ellis's behavioral characteristics. Information Research, 12 (3), paper 318. Retrieved March 25, 2010, from http://informationr.net/ir/12-3/paper318.html

Burkhardt, P. (2009). Social software trends in business: Introduction. In P. C. Deans (Ed.), Social software and Web 2.0 technology trends (pp. 1-17). Hershey, PA: Information Science Reference.

Byrne, M. D., John, N. S., Wehrle, N. S., & Crow, D. C. (1999). The tangled Web we wove: a taskonomy of WWW use. In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, 544-551.

Callery, A. (1996). Yahoo! Cataloging the Web. Retrieved March 25, 2010, from http://www.library.ucsb.edu/untangle/callery.html

Catledge, L. D., & Pitkow, J. E. (1995). Characterizing browsing strategies in the World-Wide Web. Computer Networks and ISDN Systems, 27 (6), 1065-1073.

Chan, L. M. (2007). Cataloging and classification: an introduction. The Scarecrow Press, Inc.

Chatterjee, P., Hoffman, D. L., & Novak, T. P. (2003). Modeling the clickstream: Implications for Web-based advertising efforts. Marketing Science, 22 (4), 520-541.

Chau, M., Fang, X., & Yang, C. C. (2007). Web searching in Chinese: A study of a search engine in Hong Kong. Journal of the American Society for Information Science and Technology, 58 (7), 1044-1054.

Chi, E. H., and Mytkowicz, T. (2006). Understanding navigability of social tagging systems. Retrieved March 25, 2010, from http://www.viktoria.se/altchi/submissions/submission_edchi_0.pdf

Choo C., Detlor, B., & Turnbull, D. (1999). Information seeking on the Web: An integrated model of browsing and searching. In Proceedings of the 1999 ASIS Annual Meeting. Retrieved March 25, 2010, from http://www.eric.ed.gov/ERICDocs/data/ericdocs2sql/content_storage_01/0000019b/80/16/15/6c.pdf

Choo, C., & Marton, C. (2003). Information seeking on the Web by women in IT professions. Internet Research: Electronic Networking Applications and Policy, 13 (4), 267-280.

Cockburn, A., & Jones, S. (1996). Which way now? Analysing and easing inadequacies in WWW navigation. International Journal of Human-Computer Studies, 45(1), 105–129.

Cooley, R., Mobasher, B., & Srivastava, J. (1999). Data preparation for mining World Wide Web browsing patterns. Knowledge and Information Systems.

Creswell, J. W. (2009). Research design: Qualitative, quantitative, and mixed methods approaches. Sage Publications, 1 (1), 5-32.

Creswell, J. W., & Plano Clark, V. L. (2007). Designing and conducting mixed methods research. Sage Publications.

Creswell, J. W., Plano Clark, V. L., Gutmann, M. L., & Hanson, W. E. (2003). Advanced mixed methods research designs. In A. Tashakkori, & C. Teddlie C. (Ed.), Handbook of mixed methods in social & behavioral research (pp. 209-240). Sage Publications.

Cross, R., & Parker, A. (2004). The hidden power of social networks: Understanding how work really gets done in organization. Boston, MA: Harvard Business School Press.

Dervin, B. (1992). From the mind's eye of the user: The sense-making qualitative-quantitative methodology. In J. D. Glazier, & R. R. Powell (Ed.), Qualitative research in information management (pp. 61-84). Englewood, CO: Libraries Unlimited.

Devadason, F. J., & Lingam, P. P. (1996). A methodology for the identification of information needs of users. Retrieved March 25, 2010, from http://www.ifla.org/IV/ifla62/62-devf.htm

DiNucci, D. (1999). Fragmented future. Retrieved March 25, 2010, from http://www.cdinucci.com/Darcy2/articles/Print/Printarticle7.html

Dömel, P. (1994). WebMap - A Graphical Hypertext Navigation Tool. In Proceedings of The Second International WWW Conference.

Edmunds, H. (1999). The focus group research handbook. McGraw-Hill.

Ellis, D. (1993). Modeling the information seeking patterns of academic researchers: A grounded theory approach. Library Quarterly, 63 (4), 469-486.

Ellis, D., Cox, D., & Hall, K. (1993). A comparison of the information seeking patterns of researchers in the physical and social sciences. Journal of Documentation, 49 (4), 356-369.

Erdelez, S. (1997). Information encountering: A conceptual framework for accidental information discovery. In Proceedings of an international conference on Information seeking in context, 412-421.

Erdelez, S. (1999). Information encountering: It's more than just bumping into information. Bulletin of the American Society for Information Science, February/March, 25-29.

Erdelez, S. (2004). Investigation of information encountering in the controlled research environment. Information Processing and Management, 40 (6), 1013-1025.

Farkas, M. G. (2007). Social software in libraries: Building collaboration, communication, and community online. Medford, NJ: Information Today, Inc.

Fenichel, C. H. (1981). Online searching: Measures that discriminate among users with different types of experiences. Journal of the American Society for Information Science, 32 (1), 23-32.

Ferrini, A., & Mohr, J. J. (2008). Uses, limitations, and trends in web analytics. In B. J. Jansen, A. Spink, & I. Taksa (Ed.), Handbook of research on Web log analysis (pp. 122-140). IGI Global.

Few, S. (2006). Multivariate analysis using parallel coordinates. Retrieved October 29, 2010, from http://www.perceptualedge.com/articles/b-eye/parallel_coordinates.pdf

Fink, A. (2002). The survey handbook. Thousand Oaks, CA: Sage Publications.

Ford, N. (2004). Modeling cognitive process in information seeking: From Popper to Pask. Journal of the American Society for Information Science and Technology, 55 (9), 769-782.

Furnas, G. W., Landauer, T. K., Gomez, L. M., & Dumais, S. T. (1987). The vocabulary problem in human-system communication. Communications of the ACM, 30 (11), 964-971.

Glossbrenner, E. (2001). Search engines for the World Wide Web. Pearson Education

Göker, A., & He, D. (2002). Analysing Web search logs to determine session boundaries for user-oriented learning. In Proceedings of the International Conference on Adaptive Hypermedia and Adaptive Web-Based Systems, 319–322.

Golder, S. A., & Huberman, B.A. (2006). Usage patterns of collaborative tagging systems. Journal of Information Science, 32 (2), 198-208.

Goldstein, K.M., & Blackman, S. (1978). Cognitive style: Five approaches and relevant research. New York: John Wiley.

Hammond, T., Hannay, T., Lund, B., and Scott, J. (2005). Social bookmarking tools (I). D-Lib Magazine, 11 (4). Retrieved March 25, 2010, from http://www.dlib.org/dlib/april05/hammond/04hammond.html

He, D., Göker, A., & Harper, D. J. Combining evidence for automatic Web session identification. Information Processing and Management, 38 (5), 727-742.

Hirsch, F. J., Meeks, S., & Brooks, C. L. (1997). Creating Custom Graphical Web Views Based on User Browsing History. Retrieved October 29, 2010, from http://home.comcast.net/~fjhirsch/Papers/www6/poster/paper/hg.html

Hölscher, C., & Strube, G. (2000). Web search behavior of Internet experts and newbies. Computer Network, 33, 337-346.

Hong, J. I., & Landay, J. A. (2001). WebQuilt: A framework for capturing and visualizing the Web experience. In Proceedings of The Tenth International World Wide Web Conference, 717-724.

Huang, C., Chien, L., & Oyang, Y. (2003). Relevant term suggestion in interactive web search based on contextual information in query session logs. Journal of the American Society for Information Science and Technology, 54 (7), 638-649.

Huang, C., Shen, Y., Chiang, I., & Lin, C. (2007). Characterizing web users' online information behavior. Journal of the American Society for Information Science and Technology, 58 (13), 1988-1997.

Huang, Z., Ng, J., Cheung, D. W., Ng, M. K., & Ching, W. K. (2001). A cube model and cluster analysis for Web access sessions. In Proceedings of WEBKDD 2001, 47–57.

Inselberg, A., & Dimsdale, B. (1991). Human-machine interactive system. New York: Plenum Publishing Corporation.

Jansen, B. J. (2006). Search log analysis: What it is, what's been done, how to do it. Library & Information Science Research 28, 407-432.

Jansen, B. J. (2008). The methodology of search log analysis. In B. J. Jansen, A. Spink, & I. Taksa (Ed.), Handbook of research on Web log analysis (pp. 100-123). IGI Global.

Jansen, B. J. & Pooch, U. (2001). A review of Web searching studies and a framework for future research. Journal of the American Society for Information Science, 52 (3), 235-246.

Jansen, B. J., & Spink, A. (2003). An analysis of web information seeking and use: Documents retrieved versus documents viewed. In Proceedings of the 4th International Conference on Internet computing, 65-69.

Jansen, B. J., & Spink, A. (2006). How are we searching the World Wide Web? A comparison of nine search engine transaction logs. Information Processing and Management, 42 (1), 248-263.

Jansen, B. J., Spink, A., Bateman, J., & Saracevic, T. (1998). Real life information retrieval: Study of user queries on the Web. SIGIR Forum, 32 (1), 5-17.

Jansen, B. J., Spink, A., Blakely, C., & Koshman, S. (2007). Defining a session on Web search engines. Journal of the American Society for Information Science and Technology, 58 (6), 862–871.

Jenkins, C., Corritore, C. L., Wiedenbeck, S. (2003). Patterns of information seeking on the Web: A qualitative study of domain expertise and Web expertise. IT & Society, 1 (3), 64-89.

Jiang, T., & He, D. (2007). Redefining social network services: A solution to personal information and knowledge management. In Proceedings of the 2007 IEEE/WIC/ACM International Conferences, 292-295.

Jiang, T., & Koshman, S. (2008). Understanding folksonomy as an information architecture for exploratory search. Presented at the 2008 Information Architecture Summit.

Jones, S. R., Cunningham, S. J., McNab, R., & Boddie, S. J. (2000). A transaction log analysis of a digital library. International Journal on Digital Libraries, 3 (2), 152-169.

Kalbach, J. (2007). Designing Web navigation: Optimizing the user experience. O'Reilly Media.

Kim, K. S. (2001). Information seeking on the Web: effects of user and task variables. Library and Information Science Research, 23 (3), 233-255.

Kim, K. S. (2008). Effects of emotion control and task on Web searching behavior. Information Processing and Management, 44 (1), 373-385.

Kim, K. S., & Allen, B. (2002). Cognitive and task influences on Web searching behavior. Journal of the American Society for Information Science and Technology, 52 (2), 109-119.

Kim, S. J., & Jeong, D. Y. (2006). An analysis of the development and use of theory in library and information science research articles. Library and Information Science Research, 28 (4), 548-562.

Krathwohl, D. R. (1998). Methods of Educational and Social Science Research. Waveland Press, Inc.

Krippendorff, K. (2004). Content analysis: An introduction to its methodology. Sage Publications.

Kroski, E. (2005). The hive mind: Folksonomies and user-based tagging. Retrieved March 25, 2010, from http://infotangle.blogsome.com/2005/12/07/the-hive-mind-folksonomies-and-user-based-tagging/

Krueger, R. A., & Casey, M. A. (2000). Focus groups: A practical guide for applied research. Sage Publications.

Kuhlthau, C. C. (1991). Inside the search process: Information seeking from the user's perspective. Journal of the American Society for Information Science, 42 (5), 361-371.

Lee, J., Podlaseck, M., Schonberg, E., & Hoch, R. (2001). Visualization and analysis of clickstream data of online stores for understanding web merchandising. Data Mining and Knowledge Discovery, 5, 59-84.

Marchionini, G. (1995). Information seeking in electronic environments. Cambridge University Press.

Marchionini, G., Lin, X., & Dwiggins, S. (1990). Effects of search and subject expertise on information seeking in a hypertext environment. In Proceedings of the 53rd ASIS Annual Meeting, 129-142.

Marczyk, G. R., DeMatteo, D., & Festinger, D. (2005). Essentials of research design and methodology (Essentials of behavioral science). Wiley.

Marlow, C., Naaman, M., Boyd, D., Davis, M. (2006). HT06, tagging paper, taxonomy, Flickr, academic article, ToRead. In Proceedings of the 17th Conference on Hypertext and Hypermedia, 31-40.

Martzoukou, K. (2005). A review of Web information seeking research: Considerations of method and foci of interest. Information Research, 10 (2), paper 215. Retrieved March 25, 2010, from http://informationr.net/ir/10-2/paper215.html

Marshall, C. C., & Jones, W. (2006). Keeping encountered information. Communications of the ACM, 49 (1), 66-67.

Mat-Hassan, M., & Levene (2005). Associating search and navigation behavior through log analysis. Journal of the American Society for Information Sciences and Technology, 56 (9), 913-934.

Mieszkowski, K. (2005). Steal this bookmark! Retrieved March 25, 2010, from http://www.salon.com/tech/feature/2005/02/08/tagging/

Millen, D. R., Feinberg, J., & Kerr, B. (2006). Dogear: Social bookmarking in the enterprise. In Proceedings of CHI 2006, 111-120.

Moe, W. W., & Fader, P. S. (2004). Capturing evolving visit behavior in clickstream data. Journal of Interactive Marketing, 18 (1), 5-19.

Montgomery, A. L., Li, S., Srinivasan, K., & Liechty, J. C. (2001). Modeling online browsing and path analysis using clickstream data. Retrieved March 25, 2010, from http://www.andrew.cmu.edu/user/alm3/papers/purchase%20conversion.pdf

Morgan, D. L. (1998). The focus group guidebook. Sage Publications.

Morgan, D. L., & Scannell, A. U. (1998). Planning focus groups. Sage Publications.

Morse, J. M. (1991). Approaches to qualitative-quantitative methodological triangulation. Nursing Research, 40 (2), 120-123.

Morse, J. M. (2003). Principles of mixed methods and multimethod research design. In A. Tashakkori, & C. Teddlie (Ed.), Handbook of mixed methods in social & behavioral research (pp. 189-208). Sage Publications.

Morville, P. (2006). Ambient findability: what we find changes who we become. O'Reilly.

Nahl, D. (2005). Measuring the affective information environment of Web searchers. In Proceedings of the 68th ASIS&T Annual Meeting, 41 (1), 191-197.

Nasraoui, O., Cardona, C., & Rojas, C. (2004). Mining evolving Web clickstreams with explicit retrieval similarity measures. In Proceedings of WWW 2004 International Web Dynamics Workshop. Retrieved March 25, 2010, from http://webmining.spd.louisville.edu/Websites/PAPERS/conference/Nasraoui-WebDynamics-WWW04-evolving-clickstream.pdf

Navarro-Prieto, R., Scaife, M., & Rogers, Y. (1999). Cognitive strategies in Web searching. Retrieved March 25, 2010, from http://zing.ncsl.nist.gov/hfweb/proceedings/navarro-prieto/

Newman, I., & Benz, C. R. (1998). Qualitative-quantitative research methodology: Exploring the interactive continuum. Southern Illinois University Press.

Newman, I., Ridenour, C. S., Newman, C., & DeMarco. G. (2003). A topology of research purposes and its relationship to mixed methods. In A. Tashakkori, & C. Teddlie (Ed.), Handbook of mixed methods in social & behavioral research (pp. 167-188). Sage Publications.

Newson, A., Houghton, D., & Patten, J. (2008). Blogging and other social media: Exploiting the technology and protecting the enterprise. Gower Publishing Ltd.

Noble, R. L., & Coughlin, C. (1997). Information-seeking practices of Canadian academic chemists: A study of information needs and use of resources in chemistry. Canadian Journal of Communication, 22 (3). Retrieved March 25, 2010, from http://www.cjc-online.ca/index.php/journal/article/viewArticle/1002/908

Onwuegbuzie, A. J., & Teddlie, C. (2003). A framework for analyzing data in mixed methods research. In A. Tashakkori, & C. Teddlie (Ed.), Handbook of mixed methods in social & behavioral research (pp. 351-384). Sage Publications.

O'Reilly, T. (2005). What is Web 2.0: Design patterns and business models for the next generation of software. Retrieved March 25, 2010, from http://oreilly.com/web2/archive/what-is-web-20.html

O'Reilly, T., & Battelle, J. (2009). Web squared: Web 2.0 five years on. Retrieved March 25, 2010, from http://assets.en.oreilly.com/1/event/28/web2009_websquared-whitepaper.pdf

Özmutlu, H.C., & Çavdur, F. (2005). Application of automatic topic identification on Excite Web search engine data logs. Information Processing & Management, 41 (5), 1243–1262.

Pallant, J. (2007). SPSS survival manual: a step by step guide to data analysis using SPSS for Windows. Open University Press.

Palmquist, R., & Kim, K.-S. (2000). Cognitive style and online database search experience as predictors of Web search performance. Journal of the American Society for Information Science, 51(6), 558–566.

Park, S., Bae, H., & Lee, J. (2005). End user searching: a web log analysis of NAVER, a Korean web search engine. Library and Information Science Research, 27 (2), 203-221.

Porter, J. (2008). Designing for the social web. New Riders Press.

Puchta, C., & Potter, J. (2004). Focus group practice. Sage Publications.

Quintarelli, E. (2005). Folksonomies: Power to the people. Retrieved March 25, 2010, from http://www.iskoi.org/doc/folksonomies.htm

Rainie, L., & Jansen, B. J. (2008). Surveys as a complementary method for Web log analysis. In B. J. Jansen, A. Spink, & I. Taksa (Ed.), Handbook of research on Web log analysis (pp. 39-64). IGI Global.

Rao, R. (2004). From IR to search and beyond. Queue, 2 (3), 66-73.

Rasmussen, E. M. (2003). Indexing and retrieval for the Web. Annual Review of Information Science and Technology, 37(1), 91–124.

Rieh, S. Y., & Xu, H. (2001). Patterns and sequences of multiple query reformulation in web searching: a preliminary study. In Proceedings of the 64th ASIS&T Annual Meeting, 246–255.

Ross, N., & Wolfram, D. (2000). End user searching on the internet: An analysis of term pair topics submitted to the Excite search engine. Journal of the American Society for Information Science, 51 (10), 949-958.

Rozic-Hristovski, A., Hristovski, D., & Todorovski, L. (2002). Users' information-seeking behavior on a medical library website. Journal of Medical Library Association, 90 (2), 210-217.

Schwartz, C. (2008). Thesauri and facets and tags, Oh My! A look at three decades in subject analysis. Library Trends, 56 (4), 830-842.

Sen, A., Dacin, P. A., & Pattichis, C. (2006). Current trends in Web data analysis. Communications of the ACM, 49 (11), 85-91.

Silverstein, C., Henzinger, M., Marais, H., & Moricz, M. (1998). Analysis of a very large AltaVista query log (Technical Report 1998-014). Palo Alto, CA COMPAQ System Research Center.

Sinclair, J., & Cardew-Hall, M. (2008). The folksonomy tag cloud: When is it useful? Journal of Information Science, 34 (1), 15-29.

Solomon, G., & Schrum, L. (2007). Web 2.0: New tools, new schools. International Society for Technology in Education.

Spink, A., & Cole, C. (2004). A human information behavior approach to the philosophy of information. Library Trends, 52 (3), 373-380.

Spink, A., & Jansen, B. J. (2004). Web search: Public searching of the Web. Kluwer.

Smith, G. (2008). Tagging: People-powered metadata for the social web. New Riders.

Sue, V., M., & Ritter, L. A. (2007). Conducting online survey. Sage Publications.

Sutcliffe, A. G., Ennis, M., & Watkinson, S. J. (2000). Empirical studies of end-user information searching. Journal of the American Society for Information Science, 51 (13), 1211-1231.

Svensson, M., Hk, K., Laaksolaht, J., Waern, A. (2001). Social navigation of food recipes. In Proceedings of the 2001 SIGCHI conference on human factors in computing systems, 341-348.

Tanaka, J., and Taylor, M. (1991). Object categories and expertise: is the basic level in the eye of the beholder? Cognitive Psychology, 23 (3). 457-482.

Tashakkori, A., & Tedlie, C. (1998). Mixed methodology: Combining qualitative and quantitative approaches (Applied Social Research Methods, No. 46). Sage Publications.

Tauscher, L., & Greenberg, S. (1997a). How people revisit Web pages: Empirical findings and implications for the design of history mechanisms. International Journal of Human-Computer Studies, 47(1), 97-137.

Tauscher, L., & Greenberg, S. (1997b). Revisitation Patterns in World Wide Web Navigation. In Proceedings of the 1997 Conference on Human Factors in Computing, 399-406.

Taylor, A. G. (2004). Wynar's introduction to cataloging and classification. Libraries Unlimited.

Terdiman, D. (2005). Folksonomies tap people power. Retrieved March 25, 2010, from http://www.wired.com/science/discoveries/news/2005/02/66456

Tidline, T. J. (2005). Dervin's Sense-Making. In K. E. Fisher, S. Erdelez, & L. McKechnie (Ed.), Theories of information behavior (pp. 113-117). Medford, NJ: Information Today, Inc.

Ting, I., Kimble, C., & Kudenko, D. (2004). Visualizing and classifying the pattern of user's browsing behavior for website design recommendation. In Proceedings of 1st International Workshop on Knowledge Discovery in Data Stream.

Ting, I., Clark, L., Kimble, C., Kudenko, D., & Wright, P. (2007). APD: A tool for identifying behavioral patterns automatically from clickstream data. In Proceedings of the 11th Interactional Conference on Knowledge-Based and Intelligent Information & Engineering Systems, 66-73.

Tombros, A., Ruthven, I., & Jose, J. M. (2005). How users access web pages for information seeking. Journal of the American Society for Information Science and Technology, 56 (4), 327-344.

Trant, J. (2009). Studying social tagging and folksonomy: A review and framework. Journal of Digital Information, 10 (1). Retrieved March 25, 2010, from http://journals.tdl.org/jodi/article/view/269/278

TSN Global (2008). Digital world, digital life. Retrieved October 29, 2010, from http://www.tnsglobal.com/_assets/files/TNS_Market_Research_Digital_World_Digital_Life.pdf

Wang, P., Berry, M., & Yang, Y. (2003). Mining longitudinal Web queries: Trends and patterns. Journal of the American Society for Information Science and Technology, 54 (8), 743−758.

Wang, P., Hawk, W. B., & Tenopir, C. (2000). Users' interaction with World Wide Web resources: An exploratory study using a holistic approach. Information Processing and Management, 36 (2), 229–251.

Weber, R. P. (1990). Basic content analysis. Newbury Park, CA: Sage Publications.

Wolfram, D. & Xie, H. (2000). End user database searching over the Internet: An analysis of the state of Wisconsin's BadgerLink Service. In Proceedings of the 21st National Online Meeting, 503-512.

Wolfram, D., Wang, P., & Zhang, J. (2009). Identifying Web search session patterns using cluster analysis: a comparison of three search environments. Journal of the American Society for Information Science and Technology, 60 (5), 896-910.

Williamson, K. (1998). Discovered by chance: The role of incidental information acquisition in an ecological model of information use. Library & Information Science Research, 20 (1), 23-40.

Wilson, T. D. (1981). On user studies and information needs. Journal of Documentation, 37 (1), 3-15.

Wilson, T. D. (1997). Information behavior: An interdisciplinary perspective. Information Processing and Management, 33 (4), 551-572.

Wilson, T. D. (1999). Models in information behavior research. Journal of Documentation, 55 (3), 249-270.

Wilson, T. D. (2000). Human information behavior. Information Science, 3 (2), 49-55.

Winget, M. (2006). User-defined classification on the online photo sharing site Flickr… Or, How I learned to stop worrying and love the million typing monkeys. In Proceedings of the 17th ASIS&T SIG in Classification Research Workshop.

Yang, K. (2005). Information retrieval on the Web. Annual Review of Information Science and Technology, 39(1), 33-80.

Younger, J. (2002). Metadata and libraries: What's it all about? In W. Jones, J. R. Ahronheim & J. Crwford (Ed.), Cataloging the web: Metadata, AACR, and MARC 21. Lanham, MD: The Scarecrow Press, Inc.

Zoller, A. (2007). Emerging motivations for tagging: Expression, performance and activism. In Proceedings of the 16th International World Wide Web Conference. Retrieved March 25, 2010, from http://www2007.org/workshops/paper_55.pdf