A Clickstream Data Analysis of Users' Information Seeking Modes in Social Tagging Systems

Tingting Jiang[1]
[1] School of Information Management, Wuhan University

Abstract
Social tagging systems enable their users to access useful or interesting information resources in various ways. The purposes of this study are to identify the information seeking modes adopted by users in this context and to determine the popularity as well as effectiveness of these modes. A transaction log file obtained from Douban, the most influential Chinese-language social tagging system, was examined based on an original clickstream data analysis framework. The results show that encountering, browsing by resource/tag/user/group, searching, and monitoring by user/group are the major modes ever adopted. While browsing by resource is the most popular mode, browsing by tag is the most effective one. The research findings enrich our understanding of social tagging systems as vibrant information seeking environments and provide useful implications for their interface design.
Keywords: clickstream, information seeking modes, social tagging systems, Douban
Citation: Jiang, T. (2014). A Clickstream Data Analysis of Users' Information Seeking Modes in Social Tagging Systems. In *iConference 2014 Proceedings* (p. 314–328). doi:10.9776/14091

Acknowledgements: This research has been made possible through the financial support of the Humanities and Social Sciences Research Foundation, Ministry of Education of China under grant Number 12YJC870011 and the National Natural Science Foundation of China under grant Number 71203163.
Contact: tij@whu.edu.cn

1 Introduction

The most recent revolution in the information landscape, namely Web 2.0, not only inherits the diversity and dynamism of the Web, but demonstrates even greater complexity for allowing ordinary users to create, store, and share their own information resources (Marlow et al., 2006). Accordingly, users are driven to assume the responsibilities of describing and categorizing the resources to make them findable. They achieve this through a lightweight yet efficient cataloging practice known as "tagging" – a user adding metadata or keywords to a resource (Golder & Huberman, 2006).

Tagging is essentially an individual activity since users tag according to their personal understanding and in a distributed manner. It becomes social as the social tagging system aggregates users' tags into a social classification system called "folksonomy" (Kroski, 2005). Social tagging systems, of particular interest to this study, are unconventional information systems. They are dedicated to preserving users' collections of information resources and basically rely on tagging to organize the resources (Kalbach, 2007). These two features distinguish them from other websites also supporting tagging, such as Amazon.com which has introduced customer tagging to supplement the well-constructed "departments" of products.

As more and more users register with various social tagging systems, the Web is actually experiencing the fast self-growth of numerous information repositories, many of which accommodate substantial quantities of resources. However until now these systems still have little knowledge about how their users are coping with information overload, as evidenced by the lack of relevant research. This study is among the first to investigate users' information seeking behavior in social tagging systems. To be more specific, it aims to address the following research questions:

1. What are the information seeking modes adopted by social tagging system users to find resources?

2. How popular is each mode among the users?
3. How effective is each mode in helping the users find resources worth collecting?

It is worthwhile to probe into the above questions considering that helping users find needed resources and/or discover interesting ones is among the major goals of social tagging systems (Smith, 2008). Being blind to users' actual behavior can be very dangerous to systems that live on user participation. In return for their efforts in tagging, users are expecting the expedient acquisition of needed resources. The elements associated with their frequently adopted information seeking modes, from the perspective of user-centered design, should be easily accessible on the interfaces. If users' expectations are not met, they would be less motivated to contribute tags, leading to inadequately organized systems.

2 Related Works

2.1 Theories of Information Seeking Modes

The modes in which people look for specific pieces of information have been extensively addressed in the literature on information seeking behavior. Marchionini (1995) distinguished two classes of information seeking strategies at the extremes of continua: analytical searching strategies are goal-driven and require planning; and informal browsing strategies are opportunistic and depend on interaction. According to Wilson (1997), active search, i.e. seeking out information actively, is the principal information seeking mode and complemented by three others – passive attention, passive search, and ongoing search. The two passive modes respectively refer to the unanticipated and anticipated acquisition of information, and ongoing search the update on information. In Choo et al. (1999), four scanning modes explained in a similar way, including undirected viewing, conditioned viewing, informal search, and formal search, are integrated with the behavioral model (Ellis & Haugan, 1997) describing six characteristics underlying complex information seeking patterns (starting, chaining, browsing, differentiating, monitoring, and extracting) in order to indicate which activities are likely to occur frequently for each mode.

These studies had established preliminarily the division of information seeking modes, whereas Bates (2002) provided the most focused and thorough interpretation of such division. Taking into account two dimensions – the degrees in which an individual seeks information actively and directionally, she identified searching, browsing, being aware, and monitoring as the four modes. Searching and browsing fall in the "active" category for both demanding people to invest time and effort to obtain information, but they also differ from each other because the former is in principle guided by an articulable need whereas the latter usually starts with no particular need (Bates, 2002). Correspondingly, while searchers apply cognitive resources to recall from memory certain queries that express their information needs, browsers utilize their perceptual abilities to recognize relevant information from the context (Marchionini, 1995).

Comparatively, most people are much less familiar with the being aware and monitoring modes which are often deemed informal. Being aware is simply absorbing random information that comes to us. Researchers (Erdelez, 1997; Williamson, 1998) have probed into this mode as "information encountering" in particular. Everybody encounters information, information can be encountered everywhere, and the encountered information can be used to address any purposes (Erdelez, 1999). A little different from encountering, monitoring is absorbing related information that comes to us. We do not act to find answers to the questions already in our mind but notice the answers when they appear. Social activities are very supportive of monitoring: people are likely to come across a great deal of useful information just in the process of interacting socially with others (Bates, 2002).

2.2 Information Seeking in Social Tagging Systems

Social tagging systems have grown into a promising research area (Trant, 2009). There has been a persistent interest in users' tagging behavior, including tag usage (type/subject), ranking, growth, distribution, co-

occurrence, and so forth (Golder & Huberman, 2006; Marlow et al., 2006; Kipp & Campbell, 2006; Farooq et al., 2007; Bischoff et al., 2008; Du et al., 2009; Kakali & Papatheodorou, 2010; Golbeck et al., 2011). In contrast, little work has gone specifically into users' information seeking behavior in this particular context. But we still endeavored to identify a range of existing studies that are relevant to different extents.

It has been noted that social tagging systems conduce to information exploration (Jiang & Koshman, 2008). That is, one has a good opportunity to discover unknown or unexpected resources which would not be found through directed searching (Kroski, 2005). When a user's information need is not well defined, according to Begelman et al. (2006), he or she may want to explore what other users have tagged. This is made possible by aggregating most recently or frequently tagged resources, as well as enabling pivot browsing which is a click on a username or a tag leading people to the resources collected by that user or associated with that tag (Millen, 2008). In a study on the *dogear* social bookmarking service, the results showed that approximately 60% of the visitors navigated through the aggregated collection of bookmarks by user-supplied tags, by users, or by combinations of the two (Millen & Feinberg, 2006). The findings of another study also suggested that the navigational functions of a social bookmarking service should provide sufficient information about the attached tags and social presence of other users for each bookmark (Klaisubun, et al., 2007). Such navigation is social in nature and exclusively afforded by social tagging that aims at generating a map that summarizes an explorable space (Chi & Mytkowicz, 2007). In this way, users are empowered to make new connections not predefined by the systems, allowing for innovative uses (Winget, 2006).

However known-item search in social tagging systems usually lacks effectiveness (Begelman et al., 2006). This is because folksonomies lack precision: "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). Since the vocabulary problem is inherent in free-form social tagging, the marriage of folksonomies and the controlled vocabularies used in professional indexing is advocated (Rosenfeld, 2005). Also, what should not be ignored is the problem of tag spamming caused by adding attractive yet inappropriate tags to a resource in order to draw traffic to it, which could be tackled with spam filtering and reputation mechanisms (Goh et al., 2009).

The tag cloud visualization, one of the essential socio-technical characteristics of social tagging systems, has been holding special research interest for its important role in helping users acquire resources (Trant, 2009). The tag cloud offers a visual summary of all the contents, giving users an idea of where to begin their information seeking. Scanning it requires less cognitive load than constructing search queries, especially suitable for non-specific tasks (Sinclair & Cardew-Hall, 2008). However, the typical tag cloud, where related tags are scattered as the result of the alphabetical arrangement, was challenged because meaningful connections might be missed (Hearst & Rosner, 2008). A comparative study argued that the visualization layout design relied heavily on user purposes (Lohmann et al., 2009). Continuous efforts have been made to generate thematically clustered layouts for tag clouds (Hassan-Montero & Herrero-Solana, 2006; Fujimura et al., 2008; Chen et al., 2009; Aras et al., 2010; Gou et al., 2011).

3 Method

3.1 Research Setting

Douban[1] is one of the most influential social tagging systems on the Web. A Chinese-language site founded in 2005, it has attracted more than 66 million registered users from all over the world. Douban is a social library system, to be more specific, for people to discover three types of resources – books, movies, and music albums, collect them all in one personal library, and share their libraries with others. Similar English-

[1] http://www.douban.com/

language systems include LibraryThing[2], IMDb[3], and Last.fm[4] which specialize in books, movies, and music albums respectively.

As a typical social tagging system, Douban encourages users to tag the resources in their collections and aggregates popular tags into tag clouds. Also, users are supported to meet friends and form groups in the system. Now Douban is accommodating 310,000 groups that gather users from the same geographic locations, or sharing common interests or expertise, such as "Chicago"[5], "Jazz"[6], and "Python"[7], just to name a few.

What's special about Douban is that the type of a resource determines the type of the tags assigned to it. That is, there are book tags, movie tags, and music tags, each constituting an independent folksonomy. Besides, resource collecting is made more complicated than usual. Users have to select one of the three tenses – future ("*I want to read/watch/listen to*"), present ("*I am reading/watching/listening to*"), and past perfect ("*I have read/watched/listened to*") – in order to indicate how familiar they are with the resource collected. Nevertheless, this study was conducted regardless of resource/tag types and tenses.

This study defines information seeking in Douban as looking for resources. Every time a user reaches a resource page, i.e. the page offering detailed information about the resource, one can say that she finds a resource. On the resource page, the user may perform the collecting action or just leave, signaling whether she thinks it useful or interesting. If the former, one can say that her information seeking goal is achieved.

Many social tagging systems contain primarily six categories of webpages, i.e. home page(s), resource pages, tag pages, user pages, group pages, and search pages, all of which are designed to provide access to resources. Douban is no exception:

- On home pages (general, book, movie, and music homes), users will come across unexpected resources recommended by the system, including recent, popular, and classic ones;
- Resource pages and tag pages constitute an information structure where users can make semantic navigation, i.e. accessing resources similar to current resources or associated with specific tags;
- User pages and group pages constitute a social structure where users can make social navigation, i.e. accessing resources liked by other people or groups of people;
- For users with articulable needs, resources matched with their queries will be returned on search (result) pages generated by the internal search engine.

As a whole, a vibrant information seeking environment has developed in Douban. Figure 1 demonstrates a navigation map for its information seekers. The stacks represent the above page categories and the thick arrows their hyperlinks pointing to resource pages. In addition, thinner arrows are used to indicate other available hyperlinks within each page category or across different categories. This map encompasses the major possible navigation steps, and each way they are linked up in series will engender a specific information seeking path. In particular, the resource collecting action does not belong to any of the page categories but may update the content of resource pages.

[2] http://www.librarything.com/
[3] http://www.imdb.com/
[4] http://www.last.fm/
[5] http://www.douban.com/group/chicago/
[6] http://www.douban.com/group/jazz/
[7] http://www.douban.com/group/python/

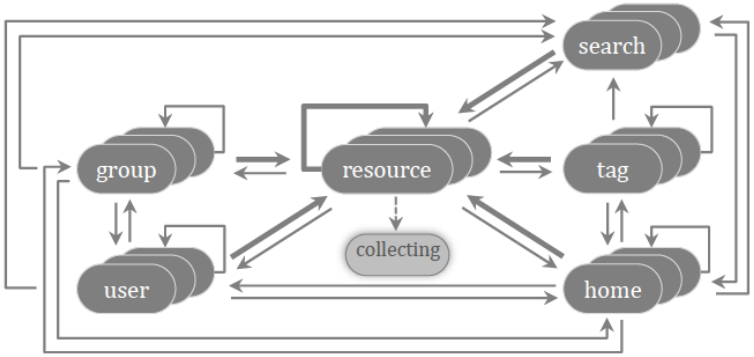

Figure 1: Hyperlinks among the major page categories in Douban

3.2 Data Collection and Cleaning

A random transaction log file was directly requested from Douban. It contains around 20 million clickstream records generated on the Web server over a 24-hour period. Websites are usually very careful about releasing transaction logs for fear of offending their users' privacy. Douban also gave full consideration to this issue and had a technician encrypt all the user identities in the log file. Specifically, each user was assigned a new ID, a string of digits that assumes no meaning but helps distinguish the user from others.

The CVS-formatted file received from Douban was imported into a single table named *original_data* in Microsoft Access. There were five basic data fields included in this table – *USER ID*, *REQUESTED_URL*, *METHOD*, *REFERRING_URL*, and *TIME*. Their descriptions are provided as follows:

- USER ID: User's IP address or username disguised with a 9 or 10-digit number that can be positive or negative;
- REQUESTED_URL: URL of the page requested by the user (the page can be visited by typing "http://www.douban.com" + "URL" in a Web browser, also applicable to the REFERRING_URL field);
- METHOD: Type of request: "GET" – requesting a page from the Web server; and "POST" – modifying the content of the data stored on the server;
- REFERRING_URL: URL of the page from which the user accesses the page in the corresponding REQUESTED_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%8	8:44:10 PM
2061537704	/subject/3189420/?i=0	GET	/subject_search?search_text=%E9%87%91%E5%AD%97%E5%A1%94%E5%8	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 2: A snippet from Table *original_data*

Figure 2 captures a snippet, comprising 24 clickstream records (or rows) belonging to 3 users, from Table *original_data* that has been sorted by *USER ID* firstly and *TIME* secondly. As can be seen, there are unrecognizable character strings starting with "%" in both the *REQUSTED_URL* and *REFERRING_URL* fields. They are actually Chinese tags or search keywords based on the UTF-8 encoding scheme. Given that this study involved no semantic analysis, they were not converted into Chinese characters.

The cleaning of Table *original_data* was completed in two steps. The first step was removing corrupted records, erroneous data produced when the Web server logged the data incorrectly. Errors can be easily detected by sorting each column in sequence because they usually appear on the top of, bottom of, or grouped together in the sorted column for not fitting the pattern of the normal data in the same column (Jansen, 2006). Next, a considerable volume of redundant records were eliminated. They failed to reflect how ordinary users navigate within Douban, e.g. requests from external sites, requests by Web search engine robots, requests for API services, and so on. Filtering such irrelevant data out helped minimize the size of the dataset and expedite the analysis.

After data cleaning, 10,303,684 clickstream records remained in the table which was then renamed *cleaned_data*. The entire *METHOD* column was deleted for displaying one invariable value – "GET", and the *USER ID*, *REQUSTED_URL*, *REFERRING_URL* fields were respectively abbreviated to *UID*, *REQ*, and *REF*. Table *cleaned_data* includes 269,658 distinct users, and 22% (N = 59,356) of them have only one 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 who has no less than 1,000 records, and the maximum number of clickstream records a user may have is 27,050.

3.3 Data Analysis

The biggest difficulty encountered in this study was that there existed no readily usable method for analyzing the above clickstream data. The popular search log analysis framework, namely, investigating search log data at the term, query, and session levels, is obviously not applicable here (Jansen, 2008). Taking into account the characteristics of clickstream data, the researcher introduced the concept "movement", defined in most dictionaries as an act of changing the location, to represent every single clickstream record in Table *cleaned_data*.

A movement describes that a certain user (*UID*) changes her location within a website, from a referring page (*REF*) to a requested page (*REQ*), at a certain time point (*TIME*). Meanwhile another concept, "footprint", was employed to refer to the requested page of a record whose referring page is in turn the footprint of the previous record. Then a movement can be represented as M_i: $F_i \leftarrow F_{i-1}$, and F_i is the footprint left as a result of M_i. Such relationship is illustrated in Figure 3.

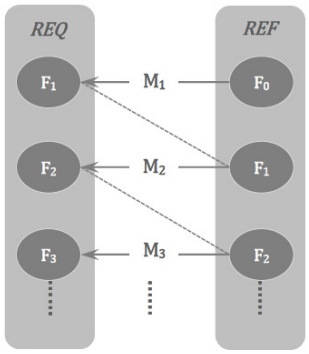

Figure 3: The relationship between footprints and movements

Footprints are visits of webpages, so the footprints left in a social tagging system also divide into six major types: F{H} (Home pages), F{R} (Resource pages), F{T} (Tag pages), F{U} (User pages), F{G} (Group pages), and F{S} (Search pages). For convenience, the following analysis deemed that the collecting action generates a seventh type of footprints, F{C}. The type of F_i determines the type of M_i. If $F_i \in F\{R\}$, M_i is a "pivotal" movement (PM) as called in this paper; if $F_i \in F\{C\}$, M_i is a "consequential" movement (CM); and otherwise, M_i is a "transitional" movement (TM).

Let's assume that a user follows the tag "interaction design" on Douban's book home to the book *Don't Make Me Think* and add it to her library. This process can be decomposed into three movements, as in Figure 4. The movement from home to tag, conducing to finding the book later, is transitional. However the PM, i.e. from tag to resource, is directly and indispensably responsible for finding the book. Collecting the book, which indicates its usefulness to the user, is the CM.

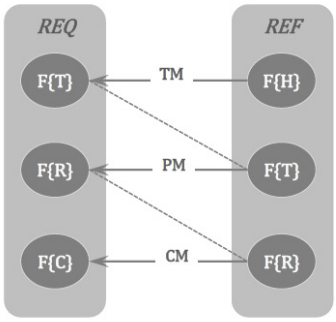

Figure 4: An illustration of transitional, pivotal, and consequential movements

PMs are critical to addressing the first research question, since the footprints one step prior to F{R} provide the most reasonable and reliable evidence regarding how users find the resources. In other words, for M_i that is a PM, the type of F_{i-1} determines the type of its characteristic information seeking mode. Therefore a new table, *pivotal_data* (Figure 5), was created by selecting all the records with the resource page URL (e.g. "*/subject/3189420/*") in the *REQ* field from Table *cleaned_data*. For example, with a search page URL displayed in the corresponding *REF* column, the first row in this table, i.e. PM: F{R}←F{S}, features the searching mode. The researcher distinguished manually all the modes ever adopted based on a thorough inspection of the entire *REF* field after sorted. The popularity of each mode was then measured with the number of all PMs featuring that mode, denoted by N_P.

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%B	7:53:24 PM
2105515094	3322741	/subject/3322741/?i=0	/subject_search?search_text=%E5%A6%82%E6%9E%9C%E4%B8	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 5: A snippet from Table *pivotal_data*

As for the third research question that concerns with the effectiveness of each mode, this study coined the "achievement rate" as a basic measure. There is an analogy between collecting a resource in a social tagging system and purchasing a product in an online retail store because they both suggest satisfaction with an item. E-commerce researchers have been using the "conversion rate", the percentage of order submissions in website visits, to measure the effectiveness of merchandising efforts (Lee et al., 2001; Ferrini & Mohr, 2008; Booth & Jansen, 2008).

As not all visits convert into purchases, not all resources found end up with being collected. That is, not all PMs are followed by CMs. Another new table, *consequential_data* (Figure 6), was created by selecting all the records with the collecting action URL (e.g. "*/j/subject/3189420/interest?interest=collect*") in the *REQ* field from Table *cleaned_data*. By jointly querying Tables *pivotal_data* and *consequential_data*, the researcher was able to tell which PMs were actually followed by CMs and counted them as effective PMs. The achievement rate of an information seeking mode was defined as the percentage of effective PMs in all PMs featuring that mode. Denoting the number of effective PMs by N_C, the achievement rate $R = N_C/N_P$.

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 6: A snippet from Table *consequential_data*

3.4 Limitations

The above research method may be limited by three major factors. First of all, the chosen research setting, Douban, is a language-specific social library system. Although it serves a remarkably large number of users, the absolute majority of them belong to the Chinese-speaking world. Both similarities and differences have been found between Web searching in Chinese and that in English (Chau et al., 2007), yet so far there is no evidence that language differences will affect users' adoption of information seeking modes. Second, the time span of the transaction log file requested from Douban is relatively short, only one day. Fortunately the considerable size of the data, exceeding 20 million records, compensated this to a certain extent. Last but not least, clickstream data analysis was the only method adopted in this research. Despite that the transaction logs provide rich unaltered information about users' behavior, they contribute little to the exploration of users' personal characteristics that may have direct influences on the ways they behave. It is hence suggested that one should introduce other methods, e.g. surveys, to tackle such shortcoming (Jansen, 2008).

4 Results

Table *pivotal_data* includes a total of 1,016,808 PMs, involved in which are 139,874 distinct users and 127,759 distinct resources. In Table *consequential_data*, the CMs add up to 239,463, involving 38,251 distinct users and 54,675 distinct resources. Therefore, among the 269,658 distinct users in Table *cleaned_data*, only 52% of them visited Douban on that day for the sake of information seeking, totally or partially, and only 27% of these information seekers eventually made some additions to their libraries.

The focused inspection of the *REF* field in Table *pivotal_data* resulted in the recognition of all major types of footprints, i.e. F{H}, F{R}, F{T}, F{U}, F{G}, and F{S}. That's to say, Douban users in reality did avail themselves of all the available access points, including home, resource, tag, user, group, and search pages, to acquire resources. These intermediaries act on resource finding in different manners, giving shape to different information seeking modes adopted by the users:

1. Encountering: F{R}←F{H};
2. Browsing by resource: F{R}←F{R};
3. Browsing by tag: F{R}←F{T};
4. Browsing by user: F{R}←F{U};
5. Browsing by group: F{R}←F{G};
6. Searching: F{R}←F{S};
7. Monitoring by user: F{R}←F{U}; and
8. Monitoring by group: F{R}←F{G}.

Searching is using the internal search engine to perform keyword search, which is the most readily understandable mode. Encountering takes place on home pages because the recommendations of resources there are made for all the people. If a resource catches a user's attention, it must happen to satisfy her interest or arouse her curiosity. And all she needs to do is an effortless click.

When browsing, in contrast, users are much more involved. They have to identify useful leads on their vague goals along the way. There are semantic leads, i.e. resources that cover specific topics and tags that describe specific topics. Meanwhile, users have their personal interests and groups common interests, and these are social leads. Thanks to the richness of hyperlinks, users are able to make pivot navigation and easily follow such leads to desired resources, achieving browsing by proxy.

Users and groups, moreover, can serve as trusted sources of monitoring. In this case, they are the users one has connected with and the groups one has affiliated to, rather than previously unknown, random ones. Keeping an eye on the updates to their resource collections may be out of socializing purposes or for information seeking. Monitoring and browsing by user/group are represented in the same form above, in

the analysis however, users or groups accessed by signed-in individuals from their own profile pages were considered as the sources of monitoring.

Figures 7 and 8 show the results obtained from analyzing the popularity and effectiveness of each information seeking mode respectively. The larger the value of N_P, the more popular a mode is. Higher achievement rate R means greater effectiveness of a mode.

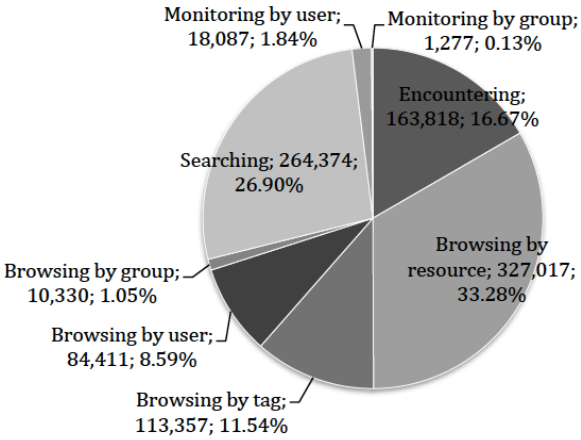

Figure 7: The popularity of different modes (N_P; proportion)

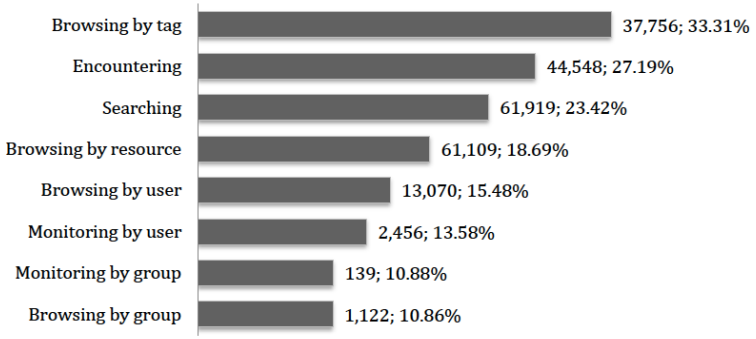

Figure 8: The effectiveness of different modes (N_C; $R = N_C/N_P$)

It is a little surprising that browsing by resource is the most popular mode, even exceeding searching. This mode takes two forms in Douban: one can browse "people like this also like" or "Doulist" for similar resources. The former is based on collaborative filtering (Linden et al., 2003), whereas the latter is a user-compiled list that contains a number of resources sharing certain attributes. The ratio of their adoption frequencies is approximately 5:1, indicating a clear preference for the former. Despite its leading popularity, the mode of browsing by resource has a poor achievement rate, even lower than the average level (22.60%). It can be inferred that the system- and human-determined similarity between resources failed to come up to users' expectations.

Another information seeking mode presenting interesting results is browsing by tag. The popularity of this mode is not competitive in Douban, which suggests general users' inadequate awareness of social tags' role in aiding exploration. But for those who attempted to obtain resources of value via tags of interest, they had a 1-in-3 chance of succeeding, making browsing by tag the most effective mode. Such finding seems

to contradict previous criticisms of folksonomies (Morville, 2005) and may to a certain extent relieve the worries about their deficiencies, especially the vocabulary problem (Golder & Huberman, 2006).

The searching and encountering modes both rank among the top three in respect of popularity and effectiveness. While searching has unarguably been the dominating mode of human's everyday online information seeking (Tombros, et al., 2005), Douban users did not depend so heavily on the internal search engine. The moderate popularity of encountering is understandable because a considerable part of what we know is absorbed this way (Bates, 2002). However unexpectedly encountering, the passive and undirected mode, is more effective than searching, the active and directed mode. This is an intriguing finding that deserves further probe.

The rest of the modes, i.e. browsing/monitoring by user/group, are all socially oriented. They are not only less frequently adopted, but also less likely to lead users to useful resources. Social tagging systems assume a dual role as information repositories and social platforms. It appears that Douban users established a clear mental boundary between the two facets, and seldom interwove information seeking with social networking activities.

5 Discussion and Conclusions

The clickstream data analysis identified eight general information seeking modes that were adopted by social tagging system users, including encountering, 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 theories of information seeking behavior (Bates, 2002), but develop in the context of social tagging systems. As a matter of fact, the universal tagging elements only contain resources, tags, and users (Smith, 2008). However, this study also took into account two functional design elements, the home and interest groups, that have become increasingly important in the architecture of social tagging systems during the past a few years.

Firstly, the home page design of the systems now thinks less of the navigational purposes and instead pays more attention to content aggregation for users' convenience. Secondly, the design of social interaction to be supported in the systems also considers 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 1.

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 home pages, the common places for everyone. These resources are usually limited and will be updated frequently. If one can find a resource of interest on the home page, therefore, it is completely opportunistic.

Although resources can be encountered elsewhere in social tagging systems, e.g. running across a resource when reading a group discussion making reference to it, they are actually ignorable compared to those encountered on the home page. As uncovered in the clickstream data analysis, encountering on home was quite popular among Douban users, accounting for 16.67% of all the resource finding occurrences, which was the third highest. The great popularity of this mode will probably be seen in other social tagging systems in that the visits to any websites usually start with the home pages. Consciously or unconsciously, users will notice the potentially interesting resources appearing there.

Meanwhile, the encountering mode was quite effective in helping users find their needed resources, with the second highest achievement rate (27.19%). But such result might be specific to Douban only. This particular system has been devoting a lot of efforts to resource recommendation and has achieved great

success. 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 home page, the effectiveness of this mode may 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, 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 her information need to a certain extent. On the contrary, encountering is viewing resources not associated with any proxy.

Among the eight information seeking modes identified, browsing by resource helped the users find 33% of the resources they ever found, which made it the most popular mode. 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 mode. It is mostly supported in social library systems, and not all of them support both forms. For example, Discogs[8] does not support the former. In spite of its popularity in Douban, this mode had an achievement rate (18.69%) even lower than the average of all the modes, suggesting unsatisfactory effectiveness. Especially, the former form will often lead users to resources that have already been viewed or collected.

In contrast, browsing by tag was the most effective mode among the eight, though only demonstrating moderate popularity. 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 characteristic information seeking mode in social tagging systems. But the clickstream data analysis showed that it was only the fourth most frequently adopted mode. Now one cannot say whether the mode is less popular in other systems too, because 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, 2005). However it was found that the achievement 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 mode failed to win overwhelming adoption, ranking the second in terms of popularity, and moreover, its achievement rate (23.42%) implies merely acceptable effectiveness. Actually this mode is mainly appropriate for tasks with specific goals. The disadvantages of Douban's search engine are 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 modes, i.e. browsing by user/group and monitoring by user/group, are all characteristic of information seeking by social proxy. This is looking 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 modes however work in

[8] http://www.discogs.com/

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, 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 modes were neither popular nor effective. They together only explained a little more than 10% of all the occurrences of resource finding and their average achievement rates (12.70%) were far below the overall average level. These may not be formal modes or they may be applicable only to users who had a passion for social activities. Social tagging systems, after all, are not social networking services such as Facebook[9] and LinkedIn[10] which connect people who are real-world acquaintances and enable them to meet new friends through the old ones. The first and foremost goal here is finding resources of interest, and the finding of interesting users or groups 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 modes.

6 References

Aras, H., Siegel, S., & Malaka, R. (2009). Semantic cloud: an enhanced browsing interface for exploring resources in folksonomy systems. In *Workshop on Visual Interfaces to the Social and Semantic Web (VISSW2010), IUI2010, Feb7, 2010, Hong Kong, China*.

Bates, M. J. (2002). Toward an integrated model of information seeking and searching. *The New Review of Information Behaviour Research*, *3*, 1-15.

Begelman, G., Keller, P., & Smadja, F. (2006, May). Automated tag clustering: Improving search and exploration in the tag space. In *Collaborative Web Tagging Workshop at WWW2006, Edinburgh, Scotland* (pp. 15-33).

Bischoff, K., Firan, C. S., Nejdl, W., & Paiu, R. (2008, October). Can all tags be used for search?. In *Proceedings of the 17th ACM conference on Information and knowledge management* (pp. 193-202). ACM.

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.

Chau, M., Xiao, F. and 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 & Technology, 58* (7), 1044-1054.

Chen, Y. X., Santamaría, R., Butz, A., & Therón, R. (2009, January). Tagclusters: Semantic aggregation of collaborative tags beyond tagclouds. In *Smart Graphics* (pp. 56-67). Springer Berlin Heidelberg.

Chi, E. H., & Mytkowicz, T. (2007, April). Understanding navigability of social tagging systems. In *Proceedings of CHI* (Vol. 7).

Choo, C. W., Detlor, B., & Turnbull, D. (1999). Information seeking on the Web: An integrated model of browsing and searching. *first monday*, *5*(2).

Du, H. S., Chu, S. K., & Lam, F. T. (2009, December). Social bookmarking and tagging behavior: an empirical analysis on delicious and connotea. In *International Conference on Knowledge Management [CD-ROM]. Hong Kong*.

Ellis, D., & Haugan, M. (1997). Modeling the information seeking patterns of engineers and research scientists in an industrial environment. *Journal of documentation*, *53*(4), 384-403.

[9] http://www.facebook.com/

[10] http://www.linkedin.com/

Erdelez, S. (1997, August). Information encountering: a conceptual framework for accidental information discovery. In *Proceedings of an international conference on Information seeking in context* (pp. 412-421). Taylor Graham Publishing.

Erdelez, S. (1999). Information encountering: It's more than just bumping into information. *Bulletin of the American Society for Information Science and Technology*, *25*(3), 26-29.

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

Farooq, U., Kannampallil, T. G., Song, Y., Ganoe, C. H., Carroll, J. M., & Giles, L. (2007, November). Evaluating tagging behavior in social bookmarking systems: metrics and design heuristics. In *Proceedings of the 2007 international ACM conference on Supporting group work* (pp. 351-360). ACM.

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.

Fujimura, K., Fujimura, S., Matsubayashi, T., Yamada, T., & Okuda, H. (2008, April). Topigraphy: visualization for large-scale tag clouds. In *Proceedings of the 17th international conference on World Wide Web* (pp. 1087-1088). ACM.

Goh, D. H. L., Chua, A., Lee, C. S., & Razikin, K. (2009). Resource discovery through social tagging: a classification and content analytic approach. *Online Information Review*, *33*(3), 568-583.

Golbeck, J., Koepfler, J., & Emmerling, B. (2011). An experimental study of social tagging behavior and image content. *Journal of the American Society for Information Science and Technology*, *62*(9), 1750-1760.

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

Gou, L., Zhang, S., Wang, J., & Zhang, X. L. (2011). Tagnet: Supporting the Exploration of Knowledge Structures of Social Tags with Multiscale Network Visualization. *International Journal of Advanced Intelligence*, *3*(1), 67-93.

Hassan-Montero, Y., & Herrero-Solana, V. (2006, October). Improving tag-clouds as visual information retrieval interfaces. In *International Conference on Multidisciplinary Information Sciences and Technologies* (pp. 25-28).

Hearst, M. A., & Rosner, D. (2008, January). Tag clouds: Data analysis tool or social signaller?. In *Hawaii International Conference on System Sciences, Proceedings of the 41st Annual* (pp. 160-160). IEEE.

Jansen, B. J. (2006). Search log analysis: What it is, what's been done, how to do it. *Library & information science research*, *28*(3), 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.

Jiang, T., & Koshman, S. (2008). Exploratory search in different information architectures. *Bulletin of the American Society for Information Science and Technology*, *34*(6), 11-13.

Kakali, C., & Papatheodorou, C. (2010). Exploitation of folksonomies in subject analysis. *Library & Information Science Research*, *32*(3), 192-202.

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

Kipp, M. E., & Campbell, D. G. (2006). Patterns and inconsistencies in collaborative tagging systems: An examination of tagging practices. *Proceedings of the American Society for Information Science and Technology*, *43*(1), 1-18.

Klaisubun, P., Kajondecha, P., & Ishikawa, T. (2007, November). Behavior patterns of information discovery in social bookmarking service. In *Web Intelligence, IEEE/WIC/ACM International Conference on* (pp. 784-787). IEEE.

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/

Lee, J., Podlaseck, M., Schonberg, E., & Hoch, R. (2001). Visualization and analysis of clickstream data of online stores for understanding web merchandising. In *Applications of Data Mining to Electronic Commerce* (pp. 59-84). Springer US.

Linden, G., Smith, B., & York, J. (2003). Amazon. com recommendations: Item-to-item collaborative filtering. *Internet Computing, IEEE*, *7*(1), 76-80.

Lohmann, S., Ziegler, J., & Tetzlaff, L. (2009). Comparison of tag cloud layouts: Task-related performance and visual exploration. In *Human-Computer Interaction–INTERACT 2009* (pp. 392-404). Springer Berlin Heidelberg.

Marchionini, G. (1995), *Information Seeking In Electronic Environments*. Cambridge University Press, New York, NY.

Marlow, C., Naaman, M., Boyd, D., & Davis, M. (2006, August). HT06, tagging paper, taxonomy, Flickr, academic article, to read. In *Proceedings of the seventeenth conference on Hypertext and hypermedia* (pp. 31-40). ACM.

McKenzie, P. J. (2003). A model of information practices in accounts of everyday-life information seeking. *Journal of documentation*, *59*(1), 19-40.

Millen, D. R. (2008). Social bookmarking and information seeking. *Information Seeking Support Systems*.

Millen, D. R., & Feinberg, J. (2006, June). Using social tagging to improve social navigation. In *Workshop on the Social Navigation and Community based Adaptation Technologies*.

Morville, P. (2005). *Ambient findability: what we find changes who we become*. O'Reilly.

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

Rosenfeld, L. (2005). Folksonomies? How about metadata ecologies? Retrieved June 5, 2011, from http://louisrosenfeld.com/home/bloug_archive/000330.html.

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

Tombros, A., Ruthven, I., & Jose, J. M. (2005). How users assess 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).

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. (1997). Information behaviour: an interdisciplinary perspective. *Information Processing & Management*, *33*(4), 551-572.

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. *Advances in Classification Research Online*, *17*(1), 1-16.

7 Table of Figures