Searching and Tagging: Two Sides of the Same Coin?

Qiaozhu Mei[†], Jing Jiang[†], Hang Su[‡], ChengXiang Zhai[†]
[†]Department of Computer Science
University of Illinois at Urbana-Champaign
[‡]Yahoo!Inc.
Santa Clara, CA

ABSTRACT

This paper presents the duality hypothesis of search and tagging, two important behaviors of web users. The hypothesis states that if a user views a document D in the search results for query Q, the user would tend to assign document D a tag identical to or similar to Q; similarly, if a user tags a document D with a tag T, the user would tend to view document D if it is in the search results obtained using T as a query. We formalize this hypothesis with a unified probabilistic model for search and tagging, and show that empirical results of several tasks on search log and tag data sets, including ad hoc search, query suggestion, and query trend analysis, all support this duality hypothesis. Since the availability of search log is limited due to the privacy concern, our study opens up a highly promising direction of using tag data to approximate or supplement search log data for studying user behavior and improving search engine accuracy.

Categories and Subject Descriptors: H.3.5 [On-line Information Services]: Web-based services

General Terms: Theory, Experimentation

Keywords: social bookmarking, query logs, tagging records, duality analysis

1. INTRODUCTION

Searching and tagging are two important activities of web users, characterizing Web 1.0 and Web 2.0, respectively. As a dominant way of accessing information, search plays important roles in our life. A search process typically involves a user typing in a keyword query, the system returning retrieved web pages, and the user clicks on result pages to view the found contents. While the goal of search is to find relevant information on the web, the query and the clickthrough information left by a user can be regarded as new information created by the user; indeed, search engine providers are collecting millions of records with queries and clickthroughs of their customers , and many techniques have been developed to leverage such log data to improve a user's searching experience through techniques such as implicit feedback [16, 23], personalized search [17, 25, 26], and query expansion/suggestion [3, 29, 6]). On the other hand, as a hallmark Web 2.0, social bookmarking systems have been attracting more and more users to tag and save their favorite URLs, and share their bookmarks online. Indeed, social bookmarking appears to be a fast growing industry. For example, Del.icio.us, a major social bookmarking website, is attracting more than a million users. Tagging typically involves a user viewing an interesting document and assigning some keyword tags to it.

Although search and tagging are apparently for different purposes—search is to find *existing* information while tagging is to create *new* information, as we will argue in this paper, they are actually very closely related, and can be regarded as two activities governed by the same common information preferences in a user's mind. In particular, both the queries posed by a user and the tags assigned by a user can be regarded as descriptions of topics interesting to the user, while the documents viewed by a user and the documents tagged by a user can both be regarded as documents relevant to the topic described by either the query or the tag. This observation can be cast as the following "duality hypothesis" of search and tagging:

Duality Hypothesis: If a user views a document D in the search results for query Q, the user would tend to assign document D a tag identical to or similar to Q; similarly, if a user tags a document D with a tag T, the user would tend to view document D if it is in the search results obtained using T as a query.

Part of the intuition behind the duality hypothesis has already been exploited in some recent work where tagging data has been exploited to improve retrieval accuracy [33]. [13] uses empirical statistics to shows that tagging data can be used to enhance search. But our formulation of the duality hypothesis explicitly reveals a more fundamental connection between search and tagging, and its potential impact goes far beyond these previous studies as will be discussed later.

The duality hypothesis is intuitively reasonable, but how can we test it more rigorously? Indeed, the hypothesis as presented above has some vague notions such as "similar" and "tend to", which make it hard to test the hypothesis. To solve this problem, we propose a unified probabilistic model for search and tagging to formalize this hypothesis. This makes it possible to test our hypothesis with empirical data and search/tagging tasks. Specifically, the model makes it possible to express search-related tasks in terms of tagging

UIUC Technical Report No: UIUCDCS-R-2007-2919

and vise versa. We show that empirical results on three tasks (i.e., ad hoc search, query suggestion, and query trend analysis) consistently support the duality hypothesis.

The duality hypothesis immediately suggests that the query log data and the tagging data can be equally valuable for inferring a user's information preferences, thus improving many information management tasks such as search and information recommendation. Although query log data has so far proven extremely useful for learning user behavior and improving search engine accuracy, the availability of search log data is unfortunately limited because of serious privacy concerns [20]. Fortunately, tag data by nature is all publicly available, and the amount of data is increasing rapidly. Thus the duality hypothesis and related probabilistic models potentially open up a highly promising new direction for using tagging data to approximate query log data to analyze user behavior and improve search accuracy.

2. DUALITY OF SEARCH AND TAGGING

2.1 Duality hypothesis

Searching and tagging have so far been considered different user behaviors on the web. Indeed, search engines and social bookmarking services provide different services to the user. Consequently, query logs and tagging records are explored in different ways. In this paper, we attempt to argue that despite the apparent differences between search and tagging, they are fundamentally connected and can be regarded as dual problems associated with the same underlying information preferences of users.

Let us first introduce the notations for a few important variables. U is a user. I is an information need. D is a web page (represented by a URL). Q is a query. L is a tag, or a bookmark used to describe a web page. Both Q and L consist of a set of words.

Our argument starts with the observation that both search and tagging can be regarded as a process to help users access information. While it is trivial to see that search is to access information, it may not be obvious why tagging or bookmarking can also be regarded as to help accessing information. On the surface, tagging is simply to mark a URL that is interesting to the user with a set of tags. However, this is not the end of this behavior. When a user finds the URL D, her information need is already satisfied. Why would she bother to tag the URL? The goal of creating a bookmark is not simply to tag the content of a URL, but to help the user access the URL next time she has the same information need or help other users to access the same information. She or other users can access the URL through the bookmarks (tags) she created. Thus the complete tagging behavior consists of information access.

This observation suggests that we can unify search and tagging on the common basis of their connections to the underlying information needs of users. Suppose a user U has an information need I and a web page D is useful in satisfying this information need. Further assume that the user would describe the information need with some keywords. The user has two ways to access page D with these keywords. One choice is to use a search engine. She would use the keyword description of the information need as a query Q, submit it to the search engine, and hope the search engine could return the URL of D. An ideal search engine will indeed return D (among other URLS) to enable the user to access D through the query Q. In reality, however, D may not show up in the search results at all. Fortunately, the user has yet another choice. She could turn to a bookmarking service, and try to access D, which would be possible if she or other users have bookmarked D with a label L that can reflect her information need I. Ideally, she has seen page D and tagged it with a label identical to the query Q, in which case, she could easily access D through the tag label L. For example, if a user tends to use the query "digital camera reviews" to search for quality reviews on cameras, she is also likely to tag a website with such reviews with the label "digital camera reviews."

Thus a query Q and the corresponding tag L can both be regarded as descriptions (or expressions) of the same information need I, and both are useful for helping users access information. Moreover, given that a query Q and a label L describe the same information need I, the associations between Q and pages viewed by a user when searching with query Q are also similar to those between L and pages tagged with L. This intuitive duality of search and tagging can be formally stated as

Duality of Searching and Tagging: Given a user $U = u$ with information need $I = i$ and some keyword expression e of i, the probability that user u views document $D = d$ when searching with query $Q = e$ is the same as the probability that u would tag document $D = d$ with tag $L = e$:

$$P(Q = e, D = d|U = u, I = s) = P(L = e, D = d|U = u, I = s),$$

or short as

$$P(Q, D|U, I) = P(L, D|U, I). \qquad (1)$$

In social tagging and collaborative search, we may reasonably assume that users who share similar information needs would behave similarly for both tagging and search. Indeed, such similarity has already been exploited in both Web search and collaborative filtering. In probabilistic terms, we can model such group behavior by assuming the joint distribution of $p(Q, D)$ (i.e., clickthrough behavior) and the joint distribution $p(L, D)$ (i.e., tagging behavior) depend only on the information need I, not the specific user U. Under this assumption we have the following *Extended Duality of Searching and Tagging*:

$$P(Q, D|I) = P(L, D|I). \qquad (2)$$

The formalization of the duality hypothesis makes it possible to test the hypothesis with empirical data and facilitates derivations of probabilistic models for solving search problems using tagging data (and vise versa). With this duality equation, we can use probability rules to derive additional equations such as $p(Q|D, I) = p(L|D, I)$, $p(Q|I) = p(L|I)$, and $p(I|Q) = p(I|L)$. We will see later that equations like these can be used to derive parallel solutions to a common problem using search log and tag log, respectively.

2.2 Unified Modeling of Search and Tagging

The duality of search and tagging suggests that it may also be possible to model search logs and tag logs in a uniform way. In this subsection, we will show that this is indeed possible, and such unified modeling provides a general way to exploit both search logs and tag logs to infer a user's information need and solve various problems such as improving search accuracy.

First, we note that the formalization of the duality hypothesis provides a basis for unifying a query Q and its equivalent tag L as one random variable E which can be interpreted as an *expression* of information need I. Indeed, using standard probabilistic rules, we can show that Equation (1) implies $p(Q = e|U, I) = p(L = e|U, I)$. We thus use E to denote either Q or L. With this unified view, we have four random variables to consider: (1) user U; (2) information need I; (3) information need expression E; and (4) document D. In general, we can observe U, E, and D, but not I; indeed, many tasks depend on inferring I based on all the other information.

Interestingly, both search logs and tag logs can be regarded as observed samples drawn from a common joint distribution $p(E, D, U)$, but with different sampling processes. Figure 1 shows the two sampling processes. Specifically, a search log record (u, e, d), where d is a document viewed by user u when searching with a query e, can be regarded as a sample obtained by first sampling a user u, then sampling an information need i (not observed), a query e, and finally a document d, whereas a tag record (u, e, d), where d is a document tagged with e by user u, can be regarded as a sample obtained by first sampling a user u, then an information need i (not observed), a document d, and finally a tag e. Furthermore, we assume that the chain of information needs of a certain user over time follows a Markov chain. These sampling processes naturally mimic the real search process and tagging process. The essence of the duality hypothesis is that while these sampling processes are different, their underlying probabilistic model is the same.

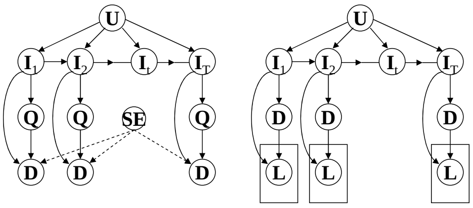

Figure 1: Two sampling processes for query logs and tagging records

Such a unified modeling of these logs has an important implication—tasks that can be accomplished by using query logs alone can also be accomplished by using tagging logs alone, and vise versa. Indeed, although query logs have been proved useful to enhance search tasks, the accessibility of query logs is highly restricted. Privacy is a severe concern that prevents query logs to be released outside search engine companies. Even within search engine companies, there is still a large debate between exploring query logs aggressively and respect the privacy of customers. On the other hand, the natural goal of a social bookmarking service is to allow users to bookmark webpages and share the bookmarks. Therefore, the tagging records are always publicly available, and there is much less privacy issue of using tagging records. Given the low accessibility of query logs on the one hand and the easy access of tagging records on the other hand, and given our duality assumption, a natural solution to problems that reply on the availability of query logs is to transform these problems into equivalent dual problems that can be solved with tagging records. The concept of *duality* is borrowed from the field of optimization: The solution of a primal problem can always be found by solving its *dual problem*, which is usually easier to deal with.

In the following section, we show that three different tasks can all be accomplished well with both query logs and tagging records, suggesting that the duality hypothesis holds. Note that while the derived models offer concrete solutions to these tasks, our goal is not to seek for optimal models or optimize the performance for these tasks, but to test the duality hypothesis through these tasks.

3. TEST THE DUALITY HYPOTHESIS

If the duality hypothesis $P(Q, D|U, I) = P(L, D|U, I)$ holds, we should expect that tasks associated with query logs can also be accomplished using the tagging records, and vise versa. In this section, we test the duality hypothesis by showing how three tasks can be accomplished using either search logs or tagging records.

To represent the search logs or tagging records we have, we organize both types of data into individual *records*, where each record consists of a user ID, a time stamp, a web page URL, and a textual expression, represented as

$$r =< u, t, d, e > .$$

The textual expression e is either a query for a search record or a tag for a tagging record, and is essentially a set of words. We assume that we either have N search records or N tagging records, represented as $\{r_i\}_{i=1}^{N}$. Let $r_i[U]$, $r_i[T]$, $r_i[D]$ and $r_i[E]$ denote the user ID, time stamp, URL and textual expression in record r_i.

The three tasks we consider are ad hoc search, query recommendation, and query trend analysis. These tasks are intentionally search-oriented rather than tagging-oriented because we want to see the feasibility of using tagging records to approximate and supplement search logs. Similar methodology can be used to examine tagging-oriented tasks. Indeed, our derivation of methods for query trend analysis can also be used to do tagging trend analysis.

3.1 Task 1: Ad Hoc Search

The task of ad hoc search can be formulated as ranking web pages based on the estimated conditional distribution $P(D|Q = e)$ given a query q. Indeed, in most work of enhancing search with query logs, this distribution, or a similar feature, is utilized [16, 30]. From the query logs, this conditional distribution can be easily estimated by

$$P(D = d|Q = e) \quad = \quad \frac{\sum_{i=1}^{N} \mathbb{I}[r_i[E] = e \wedge r_i[D] = d]}{\sum_{i=1}^{N} \mathbb{I}[r_i[E] = e]}, \qquad (3)$$

where $\mathbb{I}[S]$ is an indicator function which is set to 1 if the statement S is true and 0 otherwise.

If the duality hypothesis holds, we could assume that Q is equivalent to L, and using Equation 2, we could show that

$$P(D = d|Q = e) = P(D = d|L = e) \qquad (4)$$

This means that we can use exactly the same Equation (3) to rank web pages given a query, replacing the N search records with N tagging records.

3.2 Task 2: Query Suggestion

The task of query suggestion, or query recommendation, is to suggest alternative queries to a user based on the query she has submitted. An important criterion of query suggestion is that the suggested queries should better describe her information need. The task can then be formulated as estimating the conditional distribution $P(Q|q_0)$ given a query q_0. Based on the generative model in Figure 1, we can rewrite the probability $P(Q = e|q_0)$ as follows:

$$\begin{aligned} P(Q = e|q_0) &= \sum_s P(I = s, Q = e|q_0) \\ &= \sum_s P(Q = e|I = s)P(I = s|q_0) \qquad (5) \\ &= \sum_s P(L = e|I = s)P(I = s|q_0). \qquad (6) \end{aligned}$$

Equation (5) can be interpreted in the following way. Knowing the original query q_0 tells us something about the user's information need, and thus gives us a posterior distribution $P(I|q_0)$. With this posterior distribution, we can then obtain an adjusted distribution of Q. Given I, Q is independent of q_0. Equation 6 follows from the duality hypothesis. So once again, we obtain parallel solutions to the problem of query suggestion.

If we represent a query/tag with a bag of independent words, we can further rewrite the right side of Equation (5) and (6) as $\sum_s \prod_{w \in e} P(w|I = s)P(I = s|q_0)$.

Since we do not observe I, we cannot estimate $P(I|q_0)$ or $P(w|I)$ directly from query logs. Below we show three ways to estimate such distributions, corresponding to different assumptions about the user's change of information need. Due to the space limit, we only show the derivations for query logs; the derivation for tag logs is similar.

Unique Information Need per Record

The most strict assumption we can make here is that whenever a user moves to another query, her information need changes. In other words, we assume that there is a unique information need associated with each search record. We define $s_1, s_2, \ldots, s_N$ to be the information needs associated with each search records. Then we have

$$P(Q = e|q_0) = \sum_{i=1}^{N} \prod_{w \in e} P(w|I = s_i)P(I = s_i|q_0), \qquad (7)$$

$$where \quad P(I = s_i|q_0) \propto \mathbb{I}\big[q_0 \subseteq r_i[E]\big]. \qquad (8)$$

$P(w|I = s_i)$ can be estimated from the set of words used in the textual expression in record r_i as $\frac{\mathbb{I}[w \in r_i[E]]}{|r_i[E]|}$.

Based on the duality hypothesis, we can easily plug in

$$P(L = e|q_0) = \sum_{i=1}^{N} \prod_{w \in e} P(w|I = s_i)P(I = s_i|q_0)$$

to achieve query suggestion with tagging logs.

This assumption justifies a family of query suggestion methods which utilize information in the same records [6]. Intuitively, if the users tend to also use "spears" in their queries when they use "britney", we should recommend "britney spears" for "britney". This assumption is too strong, which only utilized the same search records. In reality, the query logs are usually sparse, and individual queries are usually short. People use two natural relaxations of this assumption as follows.

Unique Information Need per Web Page

We can relax the assumption in the first case and assume that users who clicked on the same URL to have the same information need. In this case, a URL is equivalent to a unique information need. Let M be the number of unique web pages. We then have

$$\begin{aligned} P(Q = e|q_0) &= \sum_{i=1}^{M} P(Q = e|I = s_i)P(I = s_i|q_0) \\ &= \sum_{i=1}^{M} P(Q = e|D = d_i)P(D = d_i|q_0) \qquad (9) \end{aligned}$$

Both $P(Q|D)$ and $P(D|Q)$ can be estimated from the search logs.

Parallelly, we can estimate such distribution from tagging logs:

$$P(L = e|q_0) = \sum_{i=1}^{M} P(L = e|D = d_i)P(D = d_i|q_0)$$

This relaxed assumption justifies a family of methods which uses the query-clickthrough correlation to help query suggestion [7]. Intuitively, if for URLs that are accessible by the query "svm", users also use "support vector machine" to access them, we recommend "support vector machine" to the query "svm."

Same Information Need within a Time Window

Another relaxation we can make is to assume that within a narrow time window, a user's information need does not change. We define $s_1, s_2, \ldots, s_N$ to be the information needs associated with each record in the chain of search records. We then segment this chain into $S_1, S_2, ..., S_M$, where $S_j = \{s_{t_j}, s_{t_j+1}, ..., s_{t_{j+1}-1}\}$ such that $r_{t_{j+1}}[T] - r_{t_j}[T] < \Delta t$. S_j is often referred as "query session" in literature [18]. We then have

$$P(Q = e|q_0) \propto \sum_{i=1}^{N} \sum_{j=1}^{M} P(Q = e|S_j)P(S_j|s_i)P(I = s_i|q_0),$$

$$where \quad P(S_j|s_i) = \mathbb{I}\Big[t_j \leq i < t_{j+1}\Big].$$

We compute $P(s_i|q_0)$ with Equation (8), and $P(Q = e|S_j)$ with

$$P(Q = q|S_j) = \frac{1}{(t_{j+1} - t_j)} \sum_{i=t_j}^{t_{j+1}-1} P(Q = q|I = s_i).$$

Parallelly, we can use the same assumption and utilize tagging logs to recommend query, where

$$P(L = e|q_0) \propto \sum_{i=1}^{N} \sum_{j=1}^{M} P(L = e|S_j)P(S_j|s_i)P(I = s_i|q_0).$$

This assumption justifies another family of methods which utilize query session/query chain for query suggestion. If the users tend to query "bookmark" and then refine the query as "delicious," we should recommend "delicious" to "bookmark."

3.3 Query Trend Analysis

To analyze the trend of a query over time [9], we want to estimate the function $f_q(t)$ defined as $f_q(t) = P(Q = e|T = t)$. Given a sufficient amount of search records with a wide range of time stamps, this function can be easily estimated as follows.

$$P(E = e|T = t) = \frac{\sum_{i=1}^{N} \mathbb{I}[e \subseteq r_i[E] \wedge r_i[T] = t]}{\sum_{i=1}^{N} \mathbb{I}[r_i[T] = t]}. \qquad (10)$$

Table 1: Queries used for evaluation of ad hoc search

Frequent	High Entropy	Long Tail
google	movies	data mining
yahoo	yellow pages	texas football
ebay	angelina jolie	information retrieval
		dvd camcorder

Note that with tagging records, we can use exactly the same formula to estimate the function, replacing the N search records with N tagging records.

4. EXPERIMENTS

In this section, we conduct empirical experiments to verify our hypothesis that search and tagging are dual problems. As we discussed previously, the major focus of this paper is to prove the duality hypothesis rather than to optimize a specific search task with tagging data. Analogous to duality in optimization, we show that problem A is the dual problem of B, but to find a concrete solution to optimize A is beyond the scope of this paper.

4.1 Data Collection

Query Logs: The accessibility of query logs are limited. In our experiments, we explore a query log data set released by the Microsoft Live Labs in 2006[1]. The dataset is a sample from one month's query logs collected by the MSN search engine (now Live Search[2]), containing 14.9M queries and 12.3M clicks.

Tagging Records: As a natural property of social bookmarking services, the tagging records of users are publicly accessible by others. Social bookmarking websites also provide RSS feeds (e.g. Del.icio.us RSS[3]) so that people can collect the tagging records easily. In our experiments, we collect tagging records from Del.icio.us based on the RSS feeds it provides. Note that with these RSSes, we can only access recent records of a given tag. This prevents us from collecting a complete set of tagging records. However, one can imagine a new service based on tagging records that periodically collects records through such RSSes and creates a reasonably complete tagging record dataset in a long run.

4.2 Task 1: Ad Hoc Search

To quantitatively compare the ad hoc search performance using query logs and using tagging records, we selected 10 queries, retrieved the top 20 URLs relevant to each query based on query logs and tagging records, respectively, and asked 4 human annotators to judge the relevance of these URLs on a scale of 1-5. We then measured the retrieval performance using NDCG [15]. Table 1 shows the 10 queries we chose. Table 2 shows the average NDCG measures achieved by using each type of data. This comparison shows that tagging records can indeed be used as a replacement of query logs to help ad hoc search.

Table 2: Average NDCG

query logs	0.6225
tagging records	0.7366

In Table 3, we show the top 10 URLs retrieved from each type of data for the query "yahoo." We see that the results from both types of data are highly relevant, indicating that using tagging logs to help ad hoc search is as effective as using query logs. We also see that the tagging records bring in quite a few URLs highly relevant but not in the top list by query logs (e.g. http://mail.yahoo.com/). In the last column of the table, the URLs are also ranked by tagging records, but we only include the URLs that appear in query logs. We see that the top 10 results in this column significantly overlap with the top results in the 4th column, suggesting that the users' preferences are retained if we replace query logs with tagging records.

The URL "http://www.microsoft.com/presspass/..." in the 4th column of Table 3 suggests that search results from query logs may be biased because of the search engines [17]. To further analyze this problem, we looked at the top 50 URLs ranked by query logs for "yahoo" and "lyrics," and listed some example URLs only merely relevant in Table 4. *Q-Rank* is the rank a URL receives based on query logs. For "yahoo," there are 5 URLs completely non-relevant but ranked within top 50. For "lyrics," the first two URLs are not relevant, and the last three URLs are only slightly related to "lyrics." A common characteristic of all these 10 URLs is that they are biased because the query logs are from MSN search. We found that none of these 10 URLs is ranked within top 50 by tagging records. In fact, none of them is tagged with the corresponding query, "yahoo" or "lyrics," in any tagging record. This suggests that tagging data can help filter out the bias introduced by search engines.

For certain queries, using tagging records may show advantage over using query logs. For the query "data mining," we found that we could only retrieve 3 URLs from query logs (shown as bold in Table 5). Using tagging records, however, we could get much more URLs, and the top ranked ones are all highly relevant, as shown in Table 5. These queries are generally considered to be *long tail* queries in search industry, an improvement of which could bring large benefit to the business [2].

4.3 Task 2: Query Suggestion

In the second set of experiments, we use both types of data to perform query suggestion. In Table 6, we show the top 10 suggested words for the query "yahoo," using the methods we proposed in Section 3.

First of all, most of the query words suggested by tagging records make sense, which shows that tagging data can indeed be used for query suggestion. Next, let us look at some special characteristics of the words suggested by different methods. For the three methods that use query logs, the difference among them is clear: method 1 tends to suggest different aspects of the query, such as "maps" and "finance" for "yahoo;" method 2 gives alternative ways of expressing the same query, such as "yahoo.com" and "yahoo!" for "yahoo;" method 3 other queries in the same session, and therefore suggests words that describe different but related information need, such as "google" and "aol" for "yahoo." Do recommended queries from tagging records show the same patterns? Although not as clear as shown in query logs, we can also see some similar patterns. An example is "google" as a related query word to "yahoo." Method 1 and method 2 do not suggest "google" as a related query word to

[1] http://research.microsoft.com/ur/us/fundingopps/RFPs/Search_2006_RFP_Awards.aspx

[2] http://www.live.com/

[3] http://del.icio.us/help/rss

Table 3: Comparison of the Top Search Results from Query Logs and Tagging Records

Query	Rank	from Query Logs	from Tagging Records	Filtered
yahoo	1	http://www.yahoo.com/	http://www.yahoo.com/	http://www.yahoo.com/
	2	https://login.yahoo.com/	http://my.yahoo.com/	https://login.yahoo.com/
	3	http://search.yahoo.com/	http://developer.yahoo.com/	http://finance.yahoo.com/
	4	http://finance.yahoo.com/	http://mail.yahoo.com/	http://maps.yahoo.com/
	5	http://www.messenger.yahoo.com/	https://login.yahoo.com/	http://music.yahoo.com/
	6	http://ca.yahoo.com/	http://answers.yahoo.com/	http://search.yahoo.com/
	7	http://maps.yahoo.com/	http://pipes.yahoo.com/	http://publisher.yahoo.com/
	8	http://www.microsoft.com/presspass/...	http://groups.yahoo.com/	http://ca.yahoo.com/
	9	http://personals.yahoo.com/	http://finance.yahoo.com/	http://gallery.yahoo.com/
	10	http://people.yahoo.com/	http://news.yahoo.com/	http://smallbusiness.yahoo.com/

Table 4: Biased Search Results from Query Logs

Query	Q-Rank	URL
yahoo	14	http://intl.local.live.com
	17	http://www.microsoft.com/windowsmobile/help/email.mspx
	21	http://office.microsoft.com/en-us/default.aspx
	22	http://www.microsoft.com/windows/ie/searchguide/en-en/default.mspx
	28	http://http://support.microsoft.com/?kbid=303047
lyrics	29	http://0.r.msn.com/?ld=2vbpcqlaabrz...
	30	http://0.r.msn.com/?ld=2v78qmlck6pi...
	34	http://www.microsoft.com/windows/windowsmedia/knowledgecenter/mediaadvice/0059.mspx
	35	http://www.microsoft.com/windows/windowsmedia/knowledgecenter/howto/addlyrics.aspx
	36	http://www.microsoft.com/resources/documentation/.../player_playing_files_toviewlyrics.mspx

"yahoo," but method 3 finds it because method 3 considers other tagging records within a small time window. Again, the same pattern exists with query logs. It thus confirms our assumption that a user's information need changes gradually overtime not only during searching but also during tagging.

To quantitatively compare the effectiveness of the terms suggested by query logs and by tagging records, we also conducted the following experiments. First, from the standard TREC queries 51-150, we chose 10 queries that have both query logs and tagging records associated with them in the set of query log data and tagging data we have collected. Then for each type of data (query logs or tagging records), we chose at most 10 suggested terms for each query if they appears in the TREC AP data collection. We linearly combine the suggested query model with the original query model and use the updated query model to perform regular retrieval on the AP data set. We then compare the retrieval results of using these two types of suggested terms.

As expected, utilizing the query suggestions improves the mean average precision (MAP) of retrieval, no matter whether the suggestions are from the tagging data (0.359 → 0.366) or from the search logs (0.359 → 0.364). When we combine the two types of suggestions, we further improves MAP to 0.367. Indeed, for queries such as "human genome project", we see "science", "dna", "genetic", "biology" suggested from the tagging data and "advance", "biotechnology" from search logs. Combining such suggestions improves the retrieval performance of each other.

4.4 Task 3: Query Trend Analysis

In the third set of experiments, we analyze the trends of a number of popular queries using the tagging records. Because the amount of query logs we have is only within a small time window (1 month), we use Google Trends[4] to analyze the trends of the same queries based on search volume, and compare them with the trends we have discovered. First, we show the query trends of "da vinci code" and "harry potter" between January 2005 and October 2007 in Figure 2. These two queries represent the kind of user information needs that strongly correlate with some major events. The figure on the top show the trends from tagging records, and the figure on the bottom show the trends returned from Google Trends. In the Google Trends figure, the blue line (lower line) represents "da vinci code" while the red line (upper line) represents "harry potter." As we can see, in both figures, "da vinci code" has a spike around May 2006, which corresponds to the time when the movie "The Da Vinci Code" came out. "Harry potter" has a spike around July 2007, which corresponds to the time when the latest Harry Potter movie, "Harry Potter and the Order of the Phoenix", came out. "Harry potter" also has some spikes in 2005 shown in the Google Trends figure, but only the spike around July 2005 can be seen from the tagging trend figure. This is because Del.icio.us was not popular back in 2005.

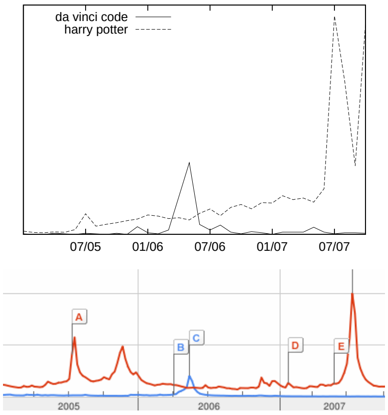

Figure 2: Long term query trends of "da vinci code" and "harry potter" between Jan 2005 and Oct 2007.

[4] http://www.google.com/trends

Table 5: Top Search Results of "Data Mining" using Tagging Records

Rank	URL	Title
1	http://datamining.typepad.com/data_mining/	Data Mining: Text Mining, Visualization and Social Media
2	http://www.anderson.ucla.edu/faculty/jason.frand...	Data Mining: What is Data Mining?
3	**http://en.wikipedia.org/wiki/Data_mining**	Data mining
4	http://www.autonlab.org/tutorials/	Statistical Data Mining Tutorials
5	http://www.jjwdesign.com/data_mining_functions.html	Data Mining Tools
6	**http://www.kdnuggets.com/**	KDnuggets: Data Mining, Web Mining, and Knowledge Discovery
7	http://www.cs.waikato.ac.nz/∼ml/weka/	Weka 3 - Data Mining Software in Java
8	http://www.kdnuggets.com/datasets/index.html	Datasets for Data Mining and Knowledge Discovery
9	http://www.applefritter.com/bannedbooks	Data Mining 101: Finding Subversives with Amazon Wishlists
10	**http://www.thearling.com/**	Data Mining and Analytic Technologies (Kurt Thearling)

Table 6: Suggested Query Words by Query Logs and Tagging Records

Query	Method	from Query Logs	from Tagging Records
yahoo	1	mail games maps finance music sbc personals my email messenger	search ajax javascript web2.0 email web news tools programming mail
	2	yahoo.com www.yahoo.com mail www.yahoo yahoo. yahoo! yaho yahoomail yah y	email search news mail portal web ajax web2.0 javascript tools
	3	google myspace ebay yahoo.com my bank free space aol news	google search web2.0 web tools blog javascript service design business

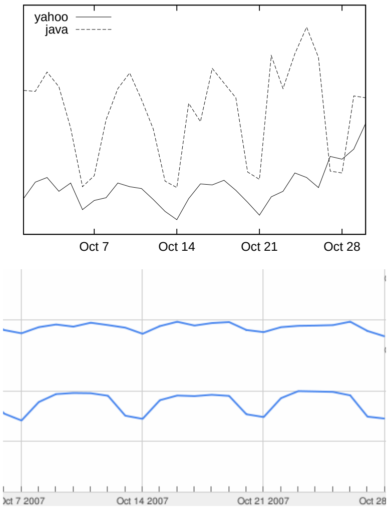

Figure 3: Weekly query trends of "yahoo" and "java" between Oct 1, 2007 and Oct 28, 2007.

Next, we show the query trends of "yahoo" and "java" between Oct 1, 2007 and Oct 28, 2007 in Figure 3. Again, the top figure is from tagging records while the bottom figure is from Google Trends. In the bottom figure, the upper line represents "yahoo" while the lower line represents "java." These two queries show very clear weekly patterns: the number of tagging records and the search volume are both lower on the weekends than on weekdays.

The similarity between the trends as shown in Figure 2 and Figure 3 again confirmed our hypothesis that searching and tagging reflect users' information need in the same way, and knowing one kind of data can help us infer the other.

5. RELATED WORK

To the best of our knowledge, there is no existing work which formally analyzes the duality between search and tagging. In most existing work on web information management, search and tagging are treated differently.

Using search engine logs (query logs) to help search has been proven to be a successful direction [24]. There has been a large body of work which utilizes query logs collected by a commercial search engine, and attempts to enhance various tasks of a search engine, e.g., search and ranking [32, 1, 7, 30, 27], personalized search [17, 25, 26], query expansion [8, 31]/query suggestion [3, 29, 6]/query substitution [18], trend analysis [9], search result organization [28], and search engine evaluation [12].

Although mining query logs has been a hot and challenging topic, the contribution from research communities outside search engine companies has so far been limited. Indeed, most work introduced above is done inside a search engine company. This is largely related to the fact that the accessibility of query logs outside the company is limited, because of severe privacy issues. On the other hand, we see that once academic researchers get the access to query logs, they can make their contributions [24, 3, 26, 28]. To alleviate this limitation, researchers use synthetic data [22], create their own search engine [23], and look for ways to alleviate the privacy concern [20]. None of this has lead to a well accepted solution so far.

One the other hand, social bookmarking is a new web service which generates large volumes of tagging logs. In recent years, researchers began to realize the importance of social bookmarking, and explored the tagging logs in different aspects [11, 21, 10, 19, 14, 33]. Most of the work focuses on the specific problems of tagging, such as folksonomy [21], tagging visualization [10] and improving the quality of tagging system [19]. [5] utilizes social bookmarking to help summarization, and [14] explores search and ranking in tagging systems. [33] is a pioneer work on using tagging logs to help web search, which utilizes tags to improve PageRank and specific types of queries (i.e., metadata queries, temporal queries, and sentiment queries). [13] uses empirical statistics to shows that tagging data can be used to enhance search. However, there is no theoretical justification for why tagging logs could help search, and they did not give a unified model to map the search task to tagging. Thus it is not clear whether tagging logs could play the role of query logs in search tasks, and whether it could help other tasks of search engines, such as query suggestion.

The duality analysis in this paper is also related to [4], which connects information filtering with information retrieval, which at that time were usually treated as independent problems.

6. CONCLUSIONS AND DISCUSSION

In this paper, we present the duality hypothesis of search and tagging. We first explain why this hypothesis intuitively makes sense, and then formalize it in probabilistic terms to facilitate testing of the hypothesis. We present two probabilistic models for search and tagging, respectively, and show how the hypothesis can be tested with three tasks that can be accomplished using either query logs or tagging records. The empirical results of the three tasks using both kinds of data support the proposed duality hypothesis.

The duality hypothesis immediately suggests that tagging data can potentially replace query log data to facilitate various services and tasks including ad hoc search, query suggestion and query trend analysis. The duality hypothesis thus provides theoretical justification for many future directions enhancing search experience with tagging records. Although currently social bookmarking services still have a biased group of users (as shown in some interesting difference between the results given by query logs and those given by tagging records), we believe that in the future, with social bookmarking becoming as popular as search itself among ordinary users, tagging records will become a much less biased source of user information that can greatly help search and other services.

In our study, users are assumed to have the same behavior and no personalization is performed due to the lack of data with user information. Thus we can only test the extended form of the duality hypothesis. It would be interesting to further test the user-specific duality hypothesis, and study how to discover a single user's preferences based on her tagging records, and use such information to build personalized services for the same users or similar users.

7. REFERENCES

[1] E. Agichtein, E. Brill, and S. Dumais. Improving web search ranking by incorporating user behavior information. In *Proceedings of SIGIR'06*, pages 19–26, 2006.

[2] C. Anderson. *The Long Tail: Why the Future of Business Is Selling Less of More.* Hyperion, 2006.

[3] D. Beeferman and A. Berger. Agglomerative clustering of a search engine query log. In *Proceedings of KDD'00*, pages 407–416, 2000.

[4] N. J. Belkin and W. B. Croft. Information filtering and information retrieval: two sides of the same coin? *Commun. ACM*, 35(12):29–38, 1992.

[5] O. Boydell and B. Smyth. From social bookmarking to social summarization: an experiment in community-based summary generation. In *Proceedings of the 12th international conference on Intelligent user interfaces*, pages 42–51, 2007.

[6] K. Church and B. Thiesson. The wild thing! In *Proceedings of the ACL 2005 on Interactive poster and demonstration sessions*, pages 93–96, 2005.

[7] N. Craswell and M. Szummer. Random walks on the click graph. In *Proceedings of SIGIR'07*, pages 239–246, 2007.

[8] H. Cui, J.-R. Wen, J.-Y. Nie, and W.-Y. Ma. Probabilistic query expansion using query logs. In *Proceedings of WWW'02*, pages 325–332, 2002.

[9] F. Diaz and R. Jones. Using temporal profiles of queries for precision prediction. In *Proceedings of SIGIR'04*, pages 18–24, 2004.

[10] M. Dubinko, R. Kumar, J. Magnani, J. Novak, P. Raghavan, and A. Tomkins. Visualizing tags over time. In *Proceedings of WWW'06*, pages 193–202, 2006.

[11] S. Golder and B. A. Huberman. The structure of collaborative tagging systems. *Journal of Information Science*, 2006.

[12] D. Hawking, N. Craswell, P. Bailey, and K. Griffihs. Measuring search engine quality. *Information Retrieval*, 4(1):33–59, 2001.

[13] P. Heymann, G. Koutrika, and H. Garcia-Molina. Can social bookmarking improve web search? In *Proceedings of the international conference on Web search and web data mining*, pages 195–206, 2008.

[14] A. Hotho, R. Jasschke, C. Schmitz, and G. Stumme. Information retrieval in folksonomies: Search and ranking. In Y. Sure and J. Domingue, editors, *The Semantic Web: Research and Applications*, volume 4011, pages 411–426, 2006.

[15] K. Jarvelin and J. Kekalainen. Ir evaluation methods for retrieving highly relevant documents. In *Proceedings of ACM SIGIR 2000*, 2000.

[16] T. Joachims. Optimizing search engines using clickthrough data. In *Proceedings of KDD'02*, pages 133–142, 2002.

[17] T. Joachims, L. Granka, B. Pan, H. Hembrooke, and G. Gay. Accurately interpreting clickthrough data as implicit feedback. In *Proceedings of SIGIR'05*, pages 154–161, 2005.

[18] R. Jones, B. Rey, O. Madani, and W. Greiner. Generating query substitutions. In *Proceedings of WWW'06*, pages 387–396, 2006.

[19] G. Koutrika, F. A. Effendi, Z. Gyöngyi, P. Heymann, and H. Garcia-Molina. Combating spam in tagging systems. In *Proceedings of the 3rd international workshop on Adversarial information retrieval on the web*, pages 57–64, 2007.

[20] R. Kumar, J. Novak, B. Pang, and A. Tomkins. On anonymizing query logs via token-based hashing. In *Proceedings of WWW'07*, pages 629–638, 2007.

[21] C. Marlow, M. Naaman, D. Boyd, and M. Davis. Position paper, tagging, taxonomy, flickr, article, toread. In *Proceedings of Hypertext*, pages 31–40, 2006.

[22] F. Radlinski and T. Joachims. Active exploration for learning rankings from clickthrough data. In *Proceedings of KDD'07*, pages 570–579, 2007.

[23] X. Shen, B. Tan, and C. Zhai. Implicit user modeling for personalized search. In *Proceedings of CIKM'05*, pages 824–831, 2005.

[24] C. Silverstein, M. R. Henzinger, H. Marais, and M. Moricz. Analysis of a very large web search engine query log. *SIGIR Forum*, 33(1):6–12, 1999.

[25] J.-T. Sun, H.-J. Zeng, H. Liu, Y. Lu, and Z. Chen. Cubesvd: a novel approach to personalized web search. In *Proceedings of WWW'05*, pages 382–390, 2005.

[26] B. Tan, X. Shen, and C. Zhai. Mining long-term search history to improve search accuracy. In *Proceedings of KDD'06*, pages 718–723, 2006.

[27] J. Teevan, E. Adar, R. Jones, and M. A. S. Potts. Information re-retrieval: repeat queries in yahoo's logs. In *Proceedings of SIGIR'07*, pages 151–158, 2007.

[28] X. Wang and C. Zhai. Learn from web search logs to organize search results. In *Proceedings of SIGIR'07*, pages 87–94, 2007.

[29] J.-R. Wen, J.-Y. Nie, and H.-J. Zhang. Clustering user queries of a search engine. In *Proceedings of WWW'01*, pages 162–168, 2001.

[30] R. W. White, M. Bilenko, and S. Cucerzan. Studying the use of popular destinations to enhance web search interaction. In *Proceedings of SIGIR'07*, pages 159–166, 2007.

[31] R. W. White and G. Marchionini. Examining the effectiveness of real-time query expansion. *Inf. Process. Manage.*, 43(3):685–704, 2007.

[32] G.-R. Xue, H.-J. Zeng, Z. Chen, Y. Yu, W.-Y. Ma, W. Xi, and W. Fan. Optimizing web search using web click-through data. In *Proceedings of CIKM'04*, pages 118–126, 2004.

[33] Y. Yanbe, A. Jatowt, S. Nakamura, and K. Tanaka. Can social bookmarking enhance search in the web? In *JCDL*, pages 107–116, 2007.