Tag-based Social Interest Discovery

Xin Li
Yahoo! Inc.
701 First Avenue
Sunnyvale, CA 94089
xl@yahoo-inc.com

Lei Guo
Yahoo! Inc.
701 First Avenue
Sunnyvale, CA 94089
lguo@yahoo-inc.com

Yihong (Eric) Zhao
Yahoo! Inc.
701 First Avenue
Sunnyvale, CA 94089
yzhao@yahoo-inc.com

ABSTRACT

The success and popularity of social network systems, such as del.icio.us, Facebook, MySpace, and YouTube, have generated many interesting and challenging problems to the research community. Among others, discovering social interests shared by groups of users is very important because it helps to connect people with common interests and encourages people to contribute and share more contents. The main challenge to solving this problem comes from the difficulty of detecting and representing the interest of the users. The existing approaches are all based on the online connections of users and so unable to identify the common interest of users who have no online connections.

In this paper, we propose a novel social interest discovery approach based on user-generated tags. Our approach is motivated by the key observation that in a social network, human users tend to use descriptive tags to annotate the contents that they are interested in. Our analysis on a large amount of real-world traces reveals that in general, user-generated tags are consistent with the web content they are attached to, while more concise and closer to the understanding and judgments of human users about the content. Thus, patterns of frequent co-occurrences of user tags can be used to characterize and capture topics of user interests. We have developed an Internet Social Interest Discovery system, *ISID*, to discover the common user interests and cluster users and their saved URLs by different interest topics. Our evaluation shows that *ISID* can effectively cluster similar documents by interest topics and discover user communities with common interests no matter if they have any online connections.

Categories and Subject Descriptors

H.3.1 [**Content Analysis and Indexing**]: Abstracting Methods; H.3.3 [**Information Search and Retrieval**]: Clustering; H.3.4 [**Systems and Software**]: Performance evaluation (efficiency and effectiveness)

General Terms

Design, Measurement, Performance

Keywords

del.icio.us, *ISID*, tag, social networks

WWW 2008 April 21–25, 2008, Beijing, China.
ACM 978-1-60558-085-2/08/04.

1. INTRODUCTION

The recent viral growth of social network systems such as del.icio.us[1], Facebook[2], MySpace[3], and YouTube[4] have created many interesting and challenging problems to the research community. In social networks, users self-organize into different communities to share the interests and contents, such as bookmarks, web blogs, questions/answers, photographs, music, and videos. Discovering common interests shared by users is a fundamental problem in social networks since it is the bread-and-butter function of building user communities of the same interests, finding the domain experts in different subjects, identifying hot social topics, and recommending personalized relevant contents. An effective and scalable solution is crucial to the growth of the social communities.

There are two kinds of existing approaches to discover shared interests in social networks. One is user-centric, which focuses on detecting social interests based on the social connections among users; the other is object-centric, which detects common interests based on the common objects fetched by users in a social community. In the user-centric approach, Schwartz *et al.,* [14] and Ali-Hasan *et al.,* [3] analyzed user's social or online connections to discover users with particular interests or expertise for a given user. Similar approach works for social networks such as Facebook. However, for social network systems such as del.icio.us, social connections among users are hard to identify. Different from this kind of approaches, we aim to find the people who share the same interests no matter whether they are connected by a social graph or not. In the object-centric approach, Sripanidkulchai *et al.,* [15] and Guo *et al.,* [9] explored the common interests among users based on the common objects they fetched in peer-to-peer networks. However, without other information of the objects, it cannot differentiate the various social interests on the same object. Furthermore, in Internet social networks such as del.icio.us, most of objects are unpopular. Thus, it is difficult to discover common interest topics of users on them. Our approach focuses on directly detecting social interests or topics by taking advantage of user tags. We cluster the related contents, i.e. URLs, and the users under the same topic. Hence, our solution removes the limitation of the object-centric approach.

In this paper, we discover common interests shared by groups of users in social networks by utilizing user tags. Our

[1]http://del.icio.us/
[2]http://www.facebook.com/
[3]http://www.myspace.com/
[4]http://www.youtube.com/

approach is based on the insightful study and observation on the user generated tags in social network systems such as del.icio.us. In these systems, people use tags as a descriptive label to annotate the content that they are interested in and to share with other users. Hence, tags implicitly and concisely represent user's interests. We have examined a large set of bookmark transactions of del.icio.us. Our extensive analysis on the data reveals the following key observations: (1) The vocabulary of all unique user tags is rich and large enough to describe the main natural concepts of the web page content of a given URL. (2) For each URL, the number of unique user tags is much smaller than the number of the unique keywords in the web page referred by the URL. The size of all user tag dictionary is much smaller than the number of unique keywords extracted from our web page corpus. (3) Different users may assign different tags to the same URL since they prefer to use personal vocabulary to summarize the same main concepts of the web page contents. However, the set of aggregated user tags on a URL is quite compact and stable enough to characterize the same main concepts of the URL. There exists a high similarity between the tag vector of a URL and the keyword vector of the URL extracted from the corresponding web page. (4) The aggregated user tags of a URL embrace different human judgments on the same subjects of the URL. This property is not possessed by the keywords of their referring web pages. Tags carrying the variation of human judgments reflects the different aspects of the same subjects. More importantly, it helps to identify the social interests in more finer granularity.

These key observations motivate us to exploit the human judgment information contained in tags to discover social interests. We have developed an Internet Social Interest Discovery system, called *ISID*, which clusters users and their saved URLs based on the user tags. Since the tags implicitly describe the users' interests, the repetitively occurrence of common tags from a set of users represent their common interests. Our evaluation results show that, (1) the URLs' contents within a *ISID* cluster have noticeably higher similarity than that of the contents of URLs across different clusters, and (2) nearly 90% of all users have their social interests discovered by the *ISID* system.

Section 2 discusses the related work. Section 3 briefly describes the real-world data traces we used for this paper. Section 4 presents the detail of our analysis as the foundation of our approach to social interest discovery. Section 5 describes *ISID* architecture to implement our approach. Section 6 presents and discusses the result of our evaluation with the real-world traces. Section 7 concludes the entire paper.

2. RELATED WORK

There have been a plenty of user-centric schemes aiming to find users with common interests. Schwartz *et al.,* [14] proposed a graph-based analysis to discover users with particular interests in email communication graphs. Referral Web [11] used the co-occurrence of names with close proximity in web documents to build referral chains and reconstruct the social relationship network. Clauset *et al.,* [7] proposed a fast community finding algorithm in large networks.

Since users in a peer-to-peer system tend to self-organize into communities, Shared interests have also been used in content locating and search in peer-to-peer networks. In the works by Sripanidkulchai *et al.,* [15] and by Guo *et al.,* [9], common interests are identified based on the common objects that different users requests. These schemes are object-centric, focusing on finding desired objects from users with the same interests. However, in these approaches, the identified shared interests are non-descriptive and implicit to the users, limiting the applications of shared interests, especially for Web social networks.

Ali-Hasan and Adamic [3] studied the social relationships through links and comments in blogs. They found that few blogging interactions reflect close offline relationships, and moreover, that many online relationships were formed through blogging. The online relationship in social networks can also be used to discover shared interests among users, however, extracting such relations is non-trivial. For example, for a social bookmark system such as del.icio.us, no such relation exists.

Tagging techniques have been widely used in different social networks, such as del.icio.us and Blog systems. However, so far there have been few experimental research on retrieving user interest related information from tags in Web social network systems. Golder *et al.,* [8] found that in del.icio.us the proportion of frequencies of tags within a given site tend to stabilize with time due to the collaborative tagging by all users. Halpin *et al.,* [10] pointed out that the distribution of frequency of del.icio.us tags for popular sites follows the power law. The authors also proposed a generative model of collaborative tagging to explain how power law distribution could arise and stabilize over time. Brooks *et al.,* [6] clustered blog articles that share the same tag, and analyzed the effectiveness of tags for blog classification. They found that the average pairwise cosine similarity of articles in tag-based clusters is only a little higher than that of randomly clustered articles, while much lower than that of articles clustered with high $tf \times idf$ key words. Different from their works, ours is based on the co-occurrence of multiple tags, instead of a single tag, thus can identify shared interests and cluster similar articles more accurately.

3. DATA SET

The data used for this paper is a partial dump of the del.icio.us database representing activity during a limited period of time.[5] In del.icio.us, when a user creates a bookmark for a URL that he/she wants to remember or share with other people, the user can add tags to this bookmark to describe it. The tags can later be used for searching, sharing, and categorizing the bookmarks. Users can add their own tags to the bookmarks pointing to the same URLs independently, called *collaborative tagging*. Different from traditional subject indexing for libraries and scientific literature, which are generated by experts, tags in del.icio.us are generated by creators and consumers of the content with freely chosen keywords rather than selected in a pre-defined term dictionary.

3.1 Data Collection and Pre-Processing

In our data set, we have 4.3 million tagged bookmarks saved by 0.2 million users on 1.4 million URLs. In order to analyze the content of these bookmarked pages, we crawled

[5]All data used in this study was anonymous in nature and only publicly-saved bookmarks were used; private bookmarks saved in del.icio.us during this time were not included in the data set.

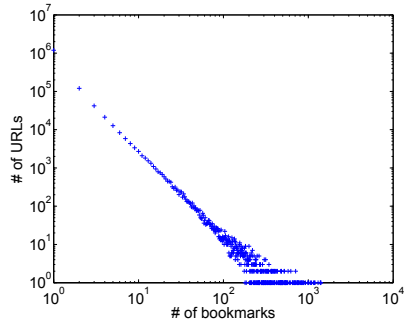


Figure 1: The distribution of the URL save frequency

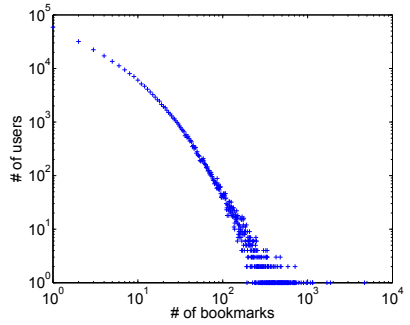


Figure 2: The distribution of the user save frequency

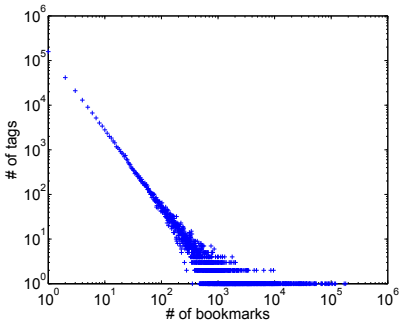


Figure 3: The distribution of the tag use frequency

the bookmarked URLs and downloaded more than 95% of the bookmarked URL pages, estimated to about 50 GB altogether. The remaining URLs were missed due to authentication requirements, dead links, unreachable sites, or server timeout. Among the downloaded pages, we discarded all non-HTML objects, *e.g.,* Adobe PDF files, MS word files, images, and flashes. We identified the language and character encoding of HTML pages, then converted them into UTF-8 encoding format and removed all non-English pages. The removed non-English HTML and non-HTML objects account for 18.3% URLs in our data set.

We used a widely used English stopword list[6] together with our own stopword dictionary to filter out all stopwords in user tags and the text of stripped HTML pages. The words left in the stripped HTML pages after filtering are called document keywords. We then normalized the remaining tags and keywords with the Porter stemming algorithm[7]. After normalization, the vocabulary of tags contains 298,350 distinct tags, while the vocabulary of document keywords contains 4,072,265 unique English words. The average number of the distinct tags (normalized) for a URL attached by all users is less than that of distinct keywords (normalized) of the same URL by the order of 100.

3.2 Users, URLs, and Tags

Figure 1 shows the distribution of the frequencies that the URLs were bookmarked in our data set. In this figure, the points are nearly in a straight line in the log-log scale, indicating that the distribution follows the power law. This observation is consistent with the Zipf-like distribution of Web object popularity [5], which shows that most Web objects are rarely accessed, while only a small number of the objects are frequently accessed.

Figure 2 shows the distribution of the bookmarking activity in our data set. The long tail of this distribution in the log-log scale means that most users are less active while a few users are highly active. As shown in the figure, most users have less than 30 bookmarks.

The long tail distribution of URL popularity and user activity have the following implications. Discovering user's common interests on Web documents is significantly different from discovering the common interests of customers in online shopping systems. It is reasonable to assume that although each individual customer may have a small number of purchases, most items should at least have a moderate number of purchases on them; otherwise these items are non-profitable. However, in a Internet social network such as del.icio.us, the distributions of user activity and URL popularity are both long-tailed: most URLs are only bookmarked once and most users only bookmark one URL. Thus, for the URLs in the tail of the popularity distribution, which account for the majority of the Web documents, it is difficult to discover common user interests, either by clustering users based on the common URLs they have bookmarked, *e.g.,* finding "similar users" that fetch same objects, or by clustering URLs bookmarked by the same user, *e.g.,* finding "similar URLs" that are fetched by the same users, like in traditional collaborative filtering approaches.

Figure 3 shows the distribution of tag frequencies in our data set. The x-axis is the number of bookmarks, and the y-axis is the number of tags with the corresponding number of bookmarks. We can see that the use of tags also follows power law distribution, meaning the selection of tags is highly concentrated. The most popular tag was used more than 180,000 times by different users altogether. We also measured the number of URLs and the number of users that can be covered by tags. Our result shows that the top popular tags connect most of the users and URLs, which motivate us to utilize tags to discover social interests among users in del.icio.us, where most users are inactive and most documents are unpopular. In the next section, we will analyze the user tags in more details.

4. ANALYSIS OF TAGS

We use the vector space model (VSM) to describe a URL. Each URL is represented with two vectors, one in the space of all tags and the other in the space of all document keywords.

In VSM, a corpus with t terms and d documents can be represented by a term-document matrix $A = (a_{ij}) \in R^{t \times d}$. Each column vector $a_j (1 \leq j \leq d)$ corresponds to a document j. Weight a_{ij} represents the importance of term i in document j. Let f_{ij} be the frequency of term i in document j. The tf-based weight of a term i in document j is

$$a_{ij}^{tf} = \frac{f_{ij}}{\sqrt{\sum_{k=1}^{t} f_{kj}^2}}. \quad (1)$$

[6]http://www.dcs.gla.ac.uk/idom/ir_resources/linguistic_utils/stop_words

[7]http://tartarus.org/ martin/PorterStemmer/

URL	http://ka1fsb.home.att.net/resolve.html
Top tf keywords	domain,name,file,resolver,server,conf,network,nameserver,ip,org,ampr
Top $tfidf$ keywords	ampr,domain,jnos,nameserver,conf,ka1fsb,resolver,ip,file,name,server
All tags	linux,howto,network,sysadmin,dns

Table 1: An example of the tf and $tf \times idf$ keywords and user-generated tags of a user-saved URL

The $tf \times idf$-based weight of a term i in document j is

$$a_{ij}^{tfidf} = \frac{b_{ij}}{\sqrt{\sum_{k=1}^{t} b_{kj}^2}}. \quad (2)$$

where b_{ij} is defined as

$$b_{ij} = f_{ij} \cdot \log(\frac{d}{D_i}), \quad (3)$$

D_i is the number of documents that contain term i, and $\log(\frac{d}{D_i})$ is called *inverse document frequency* (idf).

4.1 An Example of Tags vs. Keywords

Table 1 shows a URL bookmarked by some users, which is about the `resolv.conf` file in Linux operating systems. We show the top-10 keywords using both tf and $tf \times idf$ approaches. Along with them, we list all the tags that have been attached to this URL by all users. From this example, we can find the following properties.

First, the tags and keywords shown in Table 1 express the same content of the web page. Both tf and $tf \times idf$ keywords contain terms such "domain", "name", "file", "ip", "resolver", and so on. On the other hand, user-generated tags have a higher-level abstraction on the content. In this sense, tags and keywords both reflect the web page content, and differ only *literally*.

Second, because of its higher-level abstraction, the tags are closer to the people's understanding of the content than the keywords. For example, "sysadmin" and "dns" together carry the main purpose of file `resolv.conf`. Both tf keywords and $tf \times idf$ keywords do not have these summarization words.

Third, we can see the terms such as "ampr", "org", "jnos", "ka1fsb", and so on, which are in fact unrelated to the true purpose of this web page. They are simply used in examples and do not have direct connection with the content this web page trying to show. In other words, this web page can use any other words replacing these without changing the meaning of original content. Furthermore, these keywords will not make any sense in finding similar pages and are not useful in describing the general idea of the page. It is easy to see that the same set of tags for this page can be used to describe all other web pages with the similar content. In this sense, tags are more appropriate to describe the commonness of web pages than both tf and $tf \times idf$ keywords.

This simple, real-world example shows that intuitively, tags are more appropriate to represent human being's judgments about web content and therefore, are good candidates to represent users' interest.

4.2 The Vocabulary of Tags

Before we can use tags for capturing social interest, it is necessary to examine the vocabulary of the user-generated tags as compared with the vocabulary of keywords in the web documents. Given a web document, we are interested in seeing if the "most important" words of the document have all been covered by the vocabulary of user-generated tags. If the answer is YES, then we know that the set of user-generated tags has the comparable expression capability as the plain English words for web documents.

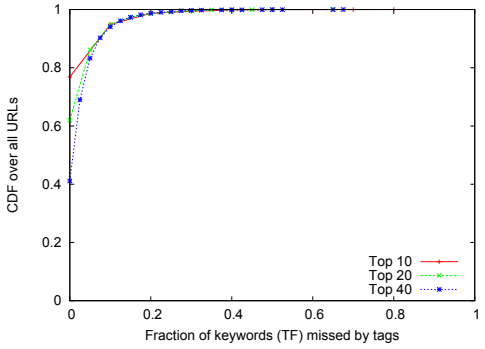


Figure 4: Tag coverage for tf keywords

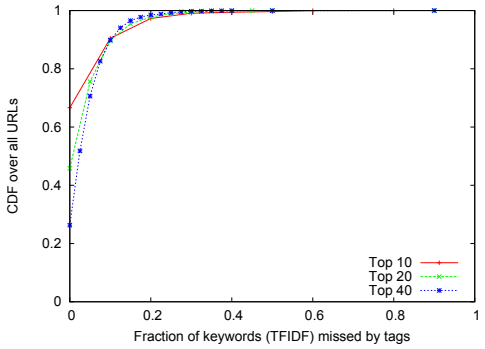


Figure 5: Tag coverage for $tf \times idf$ keywords

We measure the importance of keywords with both tf-based weight and $tf \times idf$-based weight. Figure 4 shows the coverage of user-generated tags for the tf keywords of 7000 randomly sampled English web documents in our data set. We plot the cumulative distribution function of the percentage of the missed keywords by the tag set. The three plots are for different amounts of top keywords, namely top 10, top 20, and top 40. When top 10 tf keywords chosen, 74% of all documents are fully covered by user tags. Cumulatively, the cases where the set of user tags missed at most 2 keywords accounts for 98.2% of all sampled documents. Similar conclusion can be drawn for Top-20 plot and Top-40 plot. Overall, the cases where user tags missed at most 20% of the keywords accounts for more than 98% of all documents. In fact, after careful examination, we found that most of the missed keywords are misspelled words or words invented by users, and usually cannot be found in dictionary.

Unlike the tf metric, the $tf \times idf$ metric boosts the weight of unpopular keywords, since the number of documents con-

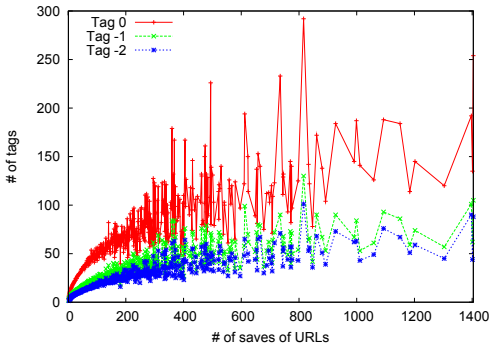


Figure 6: The convergence of tags for all URLs in our data set.

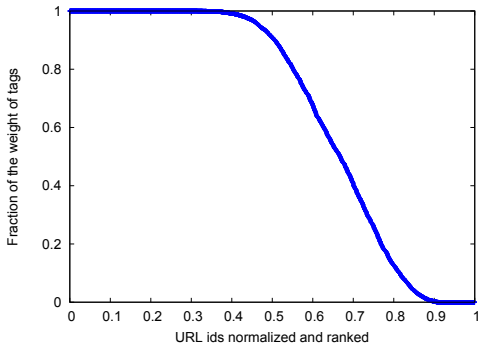


Figure 7: The distribution of tag match ratio

taining these keywords is small. However, in a large corpus, keywords with a very low document frequency are usually unpopular words, and so not good for discovering common interests or clustering similar documents. Thus, we remove the keywords whose *idf* is greater than a threshold τ, which depends on the definition of the commonness of a tag to be called to form social interest. In our study, we only consider interest topics with at least 30 URLs, and thus the threshold $\tau = \log \frac{N}{30}$. Figure 5 shows the coverage of tags on $tf{\times}idf$-based keywords for the same 7000 randomly sampled documents as above. It has similar distribution as Figure 4; for 90% of all documents, among Top-40 $tf{\times}idf$ keywords of each document, at most 10% of such top keywords cannot be covered by tags. From these two figures, we can see that the vocabulary of user-generated tags can cover the main concepts of the URLs they bookmarked.

4.3 The Convergence of User's Tag Selections

The number of distinct tags used for a given web document may increase as the document is bookmarked by more users. Golder *et al.,* [8] studied del.icio.us bookmarks and found the relative proportions of tags in the bookmarks are quite stable for popular URLs. In our study, we are more interested in the concentration and convergence of distinct tags that are used by different users. To measure the convergence of tags for all URLs, we plotted the number of distinct tags used as a function of URL popularity in Figure 6.

In this figure, the curve labeled "Tag 0" plots the function when all the tags are considered. As the popularity of URLs increases (x-axis), the number of distinct tags used for the URLs does not increases linearly. Rather, it increases with a slow speed. Notice that this plot contains "noises", those not loyal to the content of the URLs. To reduce the interference of the noisy tags, we removed all the tags used only once and plot the curve again, as shown by curve labeled "Tag -1". The consequence is the sharp reducing of the outliers, *e.g.,* at $x = 800$, and a more smooth curve. Beyond this removal of noises, further efforts will not get better convergence, as shown by curve labeled "Top -2", where all tags used less than 3 times have been removed. Therefore, the latter two curves, we believe, reflect the true convergence of the tags for all URLs. They clearly show that the total number of different tags users can use for a given document is limited no matter how popular the URL is.

4.4 Tags Matched by Documents

Now we turn to this question: How well do tags capture the main concepts of documents, or how well tags of a URL are matched by the content of the URL? Answering this question needs reviews by human editors, which cannot be covered in this paper due to page limits. In this section, we present our statistical analysis about the correlation between the tags of a URL and the content of its corresponding document. Instead of using the tf metric or $tf{\times}idf$ metric, we use the frequency of a tag in the entire corpus, *i.e.,* the total number of occurrences of this tag in our data set, as the weight to characterize the importance of this tag. The reason is as follows. For a social network system like del.icio.us, most users have the motivation to use descriptive tags for summarizing, searching, and sharing with others. So for a given set of tags for a document, the matching on a popular tag is more significant than the matching on a unpopular tag. Let $T = t_i$ be the set of tags attached to a given URL U by all the users. Let $w(t)$ be the weight of tag t, *i.e.,* the frequency of tag t in our data set. The *tag match ration* $e(T, U)$, *i.e.,* the ratio of tags of this URL that can be matched by the document is defined by the following equation

$$e(T,U) = \frac{\sum_{k|t_k \in U} w(t_k)}{\sum_i w(t_i)} \quad (4)$$

where the numerator measures the total weight of the tags that have also appeared in the keyword set of U.

The tag match ratio represents the ratio of important tags of a URL matched by the document. Figure 7 shows the distribution of tag match ratio for URLs in our data set. Each point on the x-axis represents a URL normalized by the total number of URLs. Each point on the y-axis represents the tag match ratio of a URL. For example, point $(0, 1)$ means for URL #0, all tags in its bookmarks can be matched by the document keywords of this URL, hence 100% match. As shown in this figure, the tag match ratio of nearly 50% of all URLs in our data set is one, meaning for these URLs, all tags can be covered by the corresponding documents. More than 70% of all URLs have a tag match ratio greater than 0.5, while only 10% of the URLs have no matched tags by the corresponding documents. Examination of the documents of these 10% URLs show that most of them are not the original pages, but pages used to prompt users that the original pages have been removed, or pages that require users to log in. In rare cases, users used completed unrelated tags to the page content.

4.5 Discovering Social Interest with Tags

In bookmark systems, the web pages that a user has bookmarked reflect the interest of the user. On one hand, if a user repeatedly bookmarks similar web pages, then we can say that the user have interest on the content. On the other hand, we have shown that in most cases, user-generated tags capture the content of a web page. Besides, tags are more concise and closer to the users understanding. For these reasons, we believe that tags can be used to represent the content of URLs and hence the interest of users. When multiple tags are frequently used together, they define an *topic of interest*.

It is not hard to find the frequently used tags for a given user by simple SQL-like queries. However, we are more interested in finding the sets of tags that are shared by many users on many URLs. If a set of tags are frequently used by many users, then we think that these users spontaneously form a community of interest, *even though they may not have any physical or online connections in the real world.* The tags represent the common topics of interests of these users and the URLs tagged by these users represent the commonly interested web contents to this community. Therefore, the task of discovering social interest for users is to extract frequently used tags and cluster the URLs and users under the identified tags. In a different domain of research, a question similar to ours, called *association rules*, have been explored for many years and efficient solutions have been developed. In the next Section, we will propose an architecture that uses association rules algorithms for finding frequently co-occurring tags and builds URL and user clusters for tag-based topics.

5. ARCHITECTURE FOR SOCIAL INTEREST DISCOVERY

In this section, we describe the architecture we proposed for the purpose of Internet social interest discovery (*ISID*). This architecture provides the following functions:

1. *Find topics of interests* For a given set of bookmark posts, find all topics of interests. Each topic of interests is a set of tags with the number of their co-occurrences exceeding a given threshold.
2. *Clustering* For each topic of interests, find all the URLs and the users such that those users have labeled each of the URLs with all the tags in the topic. For each topic, a user cluster and a URL cluster are generated.
3. *Indexing* Import the topics of interests and their user and URL clusters into an indexing system for application queries.

Figure 8 illustrate the software architecture of *ISID*. We discuss the detail of each *ISID* component below.

5.1 Data Source

The data source is an application data repository which stores users' posts. In general, every social network application has this data repository. The data we used for analysis in the previous sections are dumped from the data repository of del.icio.us. In addition, *ISID* requires that the data source send a stream of posts $p = (user, URL, tags)$ to *ISID*, where the combination of $user$ and URL uniquely identifies a post p. $tags$ is a set of tags labeled the URL by the $user$. This stream of posts serves as the input of *ISID*.

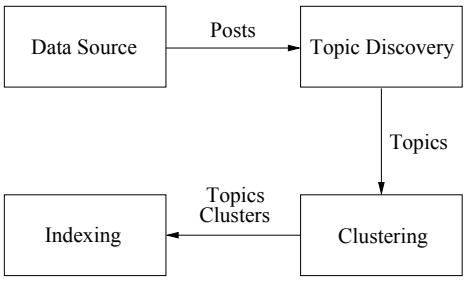


Figure 8: The software architecture of *ISID*

5.2 Topic Discovery

The function of this *ISID* component is to find the frequent tag patterns for a given set of posts[8]. The frequent pattern discovery problems have been studied in other domains. Among others, the *association rule algorithms* [1, 2] solve this problem in the domain of supermarket item-based transactions, and have been one of the top research topics for the past years. The basic idea of association rules algorithms is to discover frequent item patterns for a set of transactions and then derive the implication relationship among item sets for transactions. Consider in a supermarket, for example, a transaction may contain various *items*, such as bread and milk. If bread and milk are frequently checked out together, then the *itemset* $bread, milk$ (and all its subsets) are frequent patterns in the transactions of the supermarket. There is a parameter called *support*, which defines the threshold for an itemset to be called "frequent" if the number of transactions containing the itemset exceeds this threshold. Another important part of association rules algorithms is the ability to reduce the *implication rules* among itemsets. However, for *ISID*, we are only interested in the frequent pattern discovery part of these algorithms. We noticed that there exist other powerful approaches in the area, *e.g.,* probabilistic learning [12, 13, 16]. We decided to use the association rules algorithm for reasons of computational efficiency as well as deterministic results.

ISID uses association rules algorithms to identify the frequent tag patterns for the posts. For example, if 100 posts contain tags "food" and "recipes" while the support is 30, then the set $food, recipe$ and all its subsets are regarded as hot topics in *ISID*. That is, we have three hot topics: $\{food, recipes\}$, $\{food\}$, and $\{recipes\}$. So in *ISID*, we treat each post $p = (user, URL, tags)$ as a transaction with the key $(user, URL)$ and $tags$ (after pre-processing) as items.

Before we can build the clusters of URLs and users, we need one more step of processing. For each identified frequent itemset, an association rule algorithm produces all its subsets as derived frequent itemsets, because these subsets are also frequent. This property, however, is not always wanted in *ISID*. Consider a frequent tag pattern $\{a, b\}$ which has support $w(\{a, b\})$. If all its subsets have the same support, *i.e.,* $w(\{a\}) = w(\{b\}) = w(\{a, b\})$, we know these three different frequent patterns point to the same set of posts. In other words, they point to the same set of URLs, and represent the same common interest of user community. For removing this kind of redundancy, in *ISID*, we further

[8]At this stage, this component works offline and does not deal with streaming data.

define a topic as an tag set such that none of it subsets has the same support. Therefore, the last step of topic discovery is to compare each itemset A against each other itemset B and if $A \subseteq B$ and $w(A) = w(B)$, then remove A.

5.3 Clustering

The clustering component of *ISID* collects, for each topic (tag set), the posts that contain the tag set, and inserts the URLs and the users of the posts into two clusters. To this end, we need to scan the entire post set and match the tags of each post against all discovered topics. A naïve clustering algorithm for a given set $\mathcal{T}$ of topics and a given set $\mathcal{P}$ of posts is shown below.

```
1: for all topic T ∈ 𝒯 do
2:    T.user ← ∅
3:    T.url ← ∅
4: end for
5: for all post P ∈ 𝒫 do
6:    for all topic T of P do
7:       T.user ← T.user ∪ {P.user}
8:       T.url ← T.url ∪ {P.url}
9:    end for
10: end for
```

The most computationally intensive step here is in line 6, which matches each topic against each post. For a set of n tags, there are 2^n possible topics to check. To reduce this complexity, we build a prefix tree over the merged topics; if tag t_i and t_j are in the same topic and $w(t_i) > w(t_j)$, then t_i is an ancestor of t_j on the prefix tree. Then for each post p, we see if it contains any part of the branch on the tree. If it is, we attach the post to the ending node of the branch.

The output of this clustering algorithm is two collections of clusters identified by topics: one for URLs, where each cluster contains all the URLs that have been saved with all the tags in the topic of the cluster, and the other for users, where each cluster contains all the users who have been used all the tags in the topic of the cluster.

In addition to the URL clusters and user clusters for topics, in order to support queries based on given users and URLs, we also build the clusters for users and URLs in a similar way. For example, for each user, we build a topic cluster, which contains all the topics of the posts of the user, and a URL cluster, which contains all the URLs the of posts of the user. There is no need of extra scanning all posts, because the user-centric and URL-centric clusters can be created by embedding the statements similar to lines 7-8 within the loop at lines 6-9.

5.4 Indexing

The goal of *ISID* indexing is to provide the following basic query services for applications of *ISID*:

1. For a given topic, list all URLs that contain this topic, *i.e.,* have been tagged with all tags of the topic.
2. For a given topic, list all users that are interested in this topic, *i.e.,* have used all tags of the topic.
3. For given tags, list all topics containing the tags.
4. For a given URL, list all topics the URL belong to.
5. For a given URL and a topic, list all users that are interested in the topic and have saved the URL.

The first three queries can be resolved by indexing on topics for the topic-centric user and URL clusters. This index has to support partial match so that the topics can be found for a give set of tags. The last two queries can be resolved by indexing on the URLs for the URL-centric topic and user clusters.

6. EVALUATION RESULT

In this section, we present the result of *ISID* running on the data set described in Section 3. We first evaluate the effectiveness of *ISID* URL clusters by computing the URL similarity within and cross the clusters. Then we show how the topics discovered by *ISID* cover the individual interests of users. Finally, we present the general properties of the topic clusters.

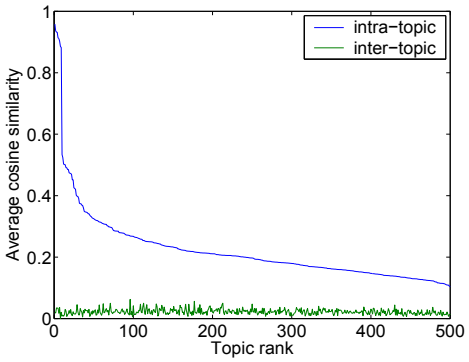
6.1 The URL Similarity of Intra- and Inter-Topics

A metric to evaluate the tag-based social interest discovery approach is whether similar contents can be well clustered under the topics, because users with shared interests are very likely to bookmark similar web pages.

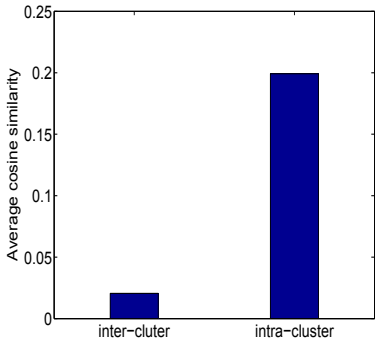
We compute the similarity between two documents with the inner product, *i.e.,* the cosine similarity, of their $tf{\times}idf$ keyword term vectors. As a comparison, we also use the tag term vector, *i.e.,* the $tf{\times}idf$ term vector of keywords that only appear in the tag vocabulary, to compute the cosine similarity of two documents. This comparison can show the effectiveness of tag vocabulary corpus in characterizing the content of tagged web pages. We randomly selected 500 interest topics, each consisting of more than 30 bookmarked URLs that share 5–6 co-occurring user tags. For each interest topic, we compute the average cosine similarity of all URL pairs in the cluster, called intra-topic similarity. We then randomly select 10,000 topic-pairs among these 500 interest topics, and compute the average pairwise document similarity between every two topics, called inter-topic similarity.

For each interest topic, we average the inter-topic similarity between this topic and all other topics among these 10,000 topic pairs, and compare it with its intra-topic similarity. Figure 9(a) shows the comparison between the intra-topic similarity and the inter-topic similarity for each interest topic in our selected topic samples, with the keyword term set. In this figure, x axis is the rank of topics, sorted by the descending order of their intra-topic similarities. y-axis shows the intra-topic similarity of each topic and the corresponding average inter-topic similarity of this topic with other topics. We can see that for all interest topics, the intra-topic similarity is consistently and significantly higher than the average inter-topic similarity with other topics. As shown in Figure 9(b), the average of intra-topic similarities across all interest topics is about 0.125, while the average of inter-topic similarities across all topic pairs is only 0.02. Figure 9(c) shows the average inter-topic similarity of topic pairs with different number of co-occurring tags. As we can see, the inter-topic similarity increases with the number of co-occurring tags. This indicates that tag co-occurrence can well cluster documents with similar content. The larger the number of co-occurring tags, the higher similar documents can be clustered.

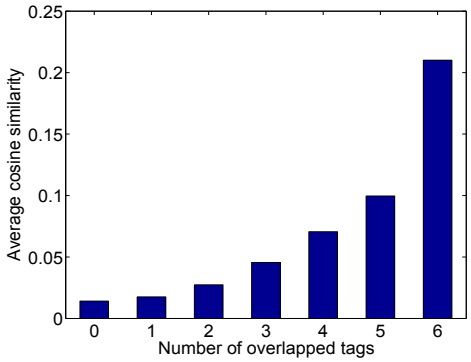
Corresponding to Figure 9, we show the comparison of the tag-based intra- and inter-topic cosine similarity for each



(a) The comparison of inter-topic and intra-topic cosine similarity

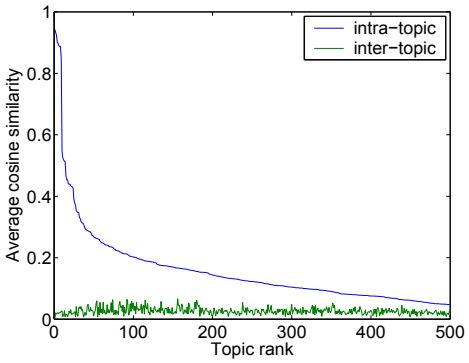


(b) Average inter- and intra-topic cosine similarity

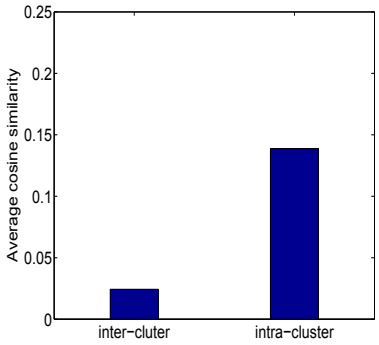


(c) Cosine similarity of topic pairs with different number of co-occurring tags

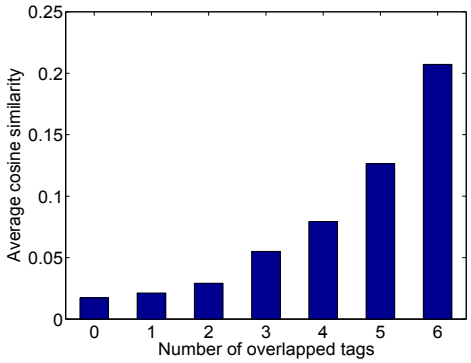
Figure 9: keyword-based cosine similarity of interest topics (support number = 30)



(a) The comparison of inter-topic and intra-topic cosine similarity



(b) Average inter- and intra-topic cosine similarity



(c) Cosine similarity of topic pairs with different number of co-occurring tags

Figure 10: Tag-based cosine similarity of interest topics (support number = 30)

interest topic, the average tag-based intra- and inter-topic similarity for all interest topics, and the average tag-based inter-topic similarity of topic pairs with different number of co-occurring tags in Figures 10(a), 10(b), and 10(c), respectively. We can see that the tag-based cosine similarity is quite close to keyword based cosine similarity, indicating that tags really capture the main concepts of documents. Thus, tags can not only be used for topic clustering, but also good enough for similarity computation. Considering that the total number of tags is only about 7.3% of the total number of keywords, tag-based topic clustering and similarity computation is not only simple and accurate, but also cost-effective in computation, because the dimension of term vector space can be significantly reduced.

Our results are significantly different from those of [4]. With the tags of blog data, Bateman *et al.,* found that the average pairwise cosine similarity of the articles in tag-based clusters is only a little higher than that of randomly clustered articles, while much lower than that of articles clustered with high $tf \times idf$ key words. However, our evaluation shows that tag-based clustering is highly accurate. The reason of this difference is that the clustering of articles in [4] is based on *single tags*, while our topic clustering is based on multiple co-occurring tags. While the common interest captured by a single tag can be very diverse, the common interest captured by a number of co-occurring tags is highly focused.

6.2 User Interest Coverage

Another important issue for *ISID* evaluation is whether the topics generated by *ISID* have indeed captured the user interests. Recall that we use tags to represent user interests. Therefore, the more frequently a user uses a tag, the higher interests he has on the corresponding topic represented by the tag. The key question to answer here is: how many of the top-used tags of each user have been captured by the topics *ISID* discovered?

For each user, we sort his tags by the number of times the tags have been used by the user. The tie is broken in favor of the tags that have been in the topics. The only reason for this is that we want our measure to have this property: for a given user, as we increase the number of his tags for consideration, from most frequently used down to most rarely used, we will see incrementally more tags that are not covered by the topics. This evaluation is to check if the top used tags of each of the users are in any topic discovered by *ISID*. Figure 11 shows the result.

In this figure, we consider three different cases: *Top-5*, where we consider only the top 5 most frequently used tags of each user and see how many of them has been in a *ISID* topic, *Top-10*, where we consider the top 10 most frequently

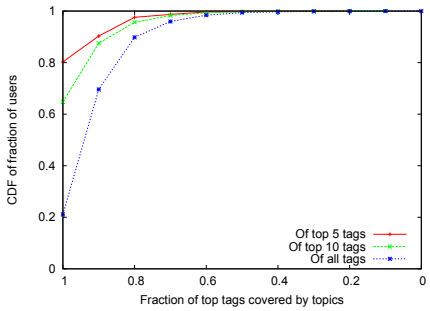


Figure 11: Topic coverage on top-frequent user tags.

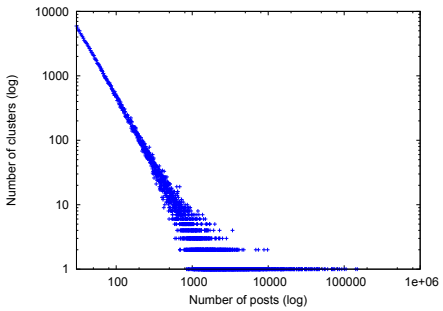


Figure 12: Number of clusters with different cluster sizes.

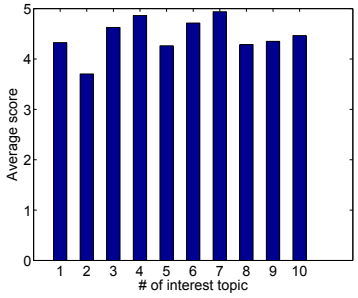


Figure 13: The results of human editorial reviews

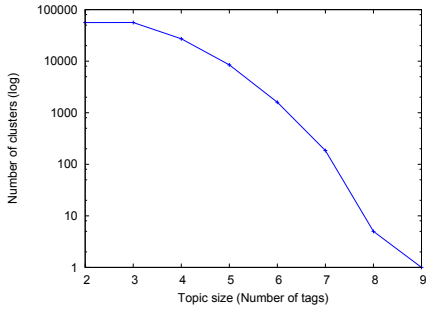


Figure 14: The distribution of the size of topics

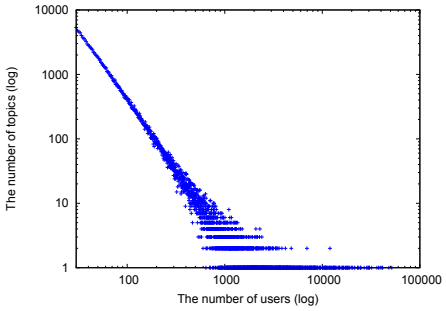


Figure 15: The distribution of topic size in terms of the number users.

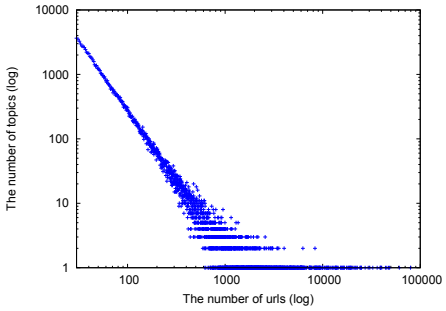


Figure 16: The distribution of topic size in terms of the number URLs.

used tags for each user, and finally *All*, where we check all the tags for each user. We can clearly see that in the *Top-5* curve, 80% of all users have all their top 5 tags in the topics and 10% of all users have 90% of their top-5 tags in the topics. Accumulatively, over 90% of all the users have at least 90% of their top-5 tags covered by *ISID* topics. When a tag is in the topic, the corresponding URLs saved with this tag and the corresponding users using this tag will go the URL and user clusters of the topic, respectively. Therefore, *ISID* has correctly identified and clustered over 90% of the interest for more than 90% of all the users. The same observation holds for the *Top-10* plot where 87% of all the users have more than 90% of their top 10 tags covered by the topics. More strikingly, even when we consider all his tags for each user, as shown by plot *All*, still 90% of all the users have more than 80% of all their tags covered by the *ISID* topics. This result shows that *the topics discovered by ISID capture the interests of users.*

6.3 Human Reviews

To evaluate the quality of the *ISID* URL clusters, we conducted a review by 4 human editors. We randomly picked 10 multi-tag topics. Within each topic, we selected the top 20 most frequently bookmarked URLs. Each of the editors need to examine all 20 URLs for each of the 10 topics. During their examination, they fill in a questionnaire about how they felt about the matching on URL contents to the topics. They are required to give the scores for each URL under a given topic. The scores are 1, 2, 3, 4, and 5, representing highly unrelated, unrelated, neutral, related, and highly related, respectively. For example, if an editor thinks a URL in a topic cluster is highly unrelated to the tags of the topic, then he will give 1 point to the URL. Figure 13 plots the average scores of the URLs for each of the 10 topics by our editorial reviews. Nine of the ten topics have an average score greater than 4. This result shows that from the human being's judgment, *ISID* indeed clusters related URLs into clusters for each topic defined by user tags.

6.4 Cluster Properties

With the support threshold 30, *ISID* generates 163 K clusters in our data set. Figure 12 shows the distribution of the number of clusters as a function of the number of URLs in the cluster. As shown in the figure, the number of clusters with a given cluster size follows an power-law distribution. The maximal cluster in our data set is 148 K, with only one topic tag "design". This plot implies that the interests of the users also follow the power-law distribution — there exists really hot topics on the Internet which capture a large amount of users, a phenomenon that has been observed for many years.

Another related question to answer is: how many tags each of the topics contains? Figure 14 plots the number of clusters as a function of the number of tags for each multi-tag topics. We can see that most of the topics have no more than 5 tags, another fact that users tend to use a small number of words to summarize the contents for themselves. Also note that beyond 6 tags, the number of clusters reduces quickly. This phenomenon implies that beyond 6 words, the users are unlikely to reach consensus about the terms for describing a given content.

To complete our result report, finally, we show the distribution of the number of topics as the function of the number of users and the number of URLs. Not surprisingly, these

two distributions also follow the power-law, as shown in Figure 15 and Figure 16.

7. CONCLUSION

In this paper, we have proposed a tag-based social interest discovery approach. We justified that user-generated tags are effective to represent user interests because these tags reflect human being's judgments while more concise and closer to human understanding. So the consensus among users for the content of a given web page can be reached more likely via tags than via keywords. We have implemented a system to discover common interest topics in social networks such as del.icio.us, without any information on the online or offline social connections among users.

8. ACKNOWLEDGMENTS

We thank Joshua Schachter and Toby Elliott for providing us feedbacks and access to the del.icio.us data base. We thank Zhichen Xu and Yun Fu for their feedbacks during the very early stage of this research work, Dave Khor for his help in building the *ISID* system, Li Xu and Tie Wang for their comments on the evaluation of the results, and Ming Sui, Yongdong Wang, and Stephen Hood for their reviews of our draft. Finally, we would like to thank all the anonymous reviewers for all their comments and suggestions.

9. REFERENCES

[1] R. Agrawal, T. Imieliński, and A. Swami. Mining association rules between sets of items in large databases. In *Proc. of ACM SIGMOD*, pages 207–216, June 1993.

[2] R. Agrawal and R. Srikant. Fast algorithms for mining association rules in large databases. In *Proc. VLDB*, pages 487–499, Sept. 1994.

[3] N. Ali-Hasan and L. Adamic. Expressing social relationships on the blog through links and comments. In *Proc. of International Conference on Weblogs and Social Media*, Mar. 2007.

[4] S. Bateman, C. Brooks, G. McCalla, and P. Brusilovsky. Applying collaborative tagging to e-learning. In *Proc. of ACM WWW*, May 2007.

[5] L. Breslau, P. Cao, L. Fan, G. Philips, and S. Shenker. Web caching and Zipf-like distributions: Evidence and implications. In *Proc. of INFOCOM*, Mar. 1999.

[6] C. H. Brooks and N. Montanez. Improved annotation of blogosphere via autotagging and hierarchical clustering. In *Proc. of ACM WWW*, pages 625–631, May 2006.

[7] A. Clauset, M. E. J. Newman, and C. Moore. Finding community structure in very large networks. *Physical Review E*, 70(066111), 2004.

[8] S. A. Golder and B. A. Huberman. Usage patterns of collaborative tagging system. *Journal of Information Science*, 32(2):198–208, 2006.

[9] L. Guo, S. Jiang, L. Xiao, and X. Zhang. Fast and low-cost search schemes by exploiting localities in p2p networks. *Journal of Parallel and Distributed Computing*, 65(6):729–742, 2005.

[10] H. Halpin, V. Robu, and H. Shepherd. The complex dynamics of collaborative tagging. In *Proc. of ACM WWW*, pages 211–220, May 2007.

[11] H. Kautz, B. Selman, and M. Shah. Referral Web: combining social networks and collaborative filtering. *Communications of the ACM*, 40(3):63–65, 1997.

[12] K. Lerman, A. Plangrasopchok, and C. Wong. Personalizing results of image search on flickr. In *AAAI workshop on Intelligent Techniques for Web Personlization*, 2007.

[13] A. Plangprasopchok and K. Lerman. Exploiting social annotation for automatic resource discovery. In *AAAI workshop on Information Integration from the Web*, 2007.

[14] M. F. Schwartz and D. C. M. Wood. Discovering shared interests using graph analysis. *Communications of the ACM*, 36(8):78–89, 1993.

[15] K. Sripanidkulchai, B. Maggs, and H. Zhang. Efficient content location using interest-based locality in peer-to-peer systems. In *Proc. of INFOCOMM*, Mar. 2003.

[16] X. Wu, L. Zhang, and Y. Yu. Exploring social annotations for the semantic web. In *WWW '06: Proceedings of the 15th international conference on World Wide Web*, pages 417–426, New York, NY, USA, 2006. ACM.