Automobile, Car and BMW: Horizontal and Hierarchical Approach in Social Tagging Systems

Tom Chao Zhou
czhou@cse.cuhk.edu.hk

Irwin King
king@cse.cuhk.edu.hk

Department of Computer Science and Engineering
The Chinese University of Hong Kong, Shatin, Hong Kong

ABSTRACT

Social tagging systems have recently emerged as an effective way for users to annotate and organize large collections of resources on the Web. Moreover, they also facilitate an efficient sharing of vast amounts of resources among different users. In this paper, we analyze tags' usage pattern in real world data sets and find that among tags representing the same concept, some tags are less popular, resulting in reduced exploring effectiveness in the current social tagging systems. Another limitation is that users cannot roll-up or drill-down the concept hierarchy of tag queries, resulting in the limited scope of service and a failure to meet users' dynamic information needs which often change with the current information provided. In order to overcome these shortcomings, we propose a novel three-phase approach as the following: (1) finding semantically-related tags for tag query; (2) constructing clusters of tags representing the same concept; and (3) building hierarchical relationships among clusters of tags. Based on our approaches, horizontal and hierarchical exploration can be implemented. Experiments employing real world dataset show encouraging results and confirm that the proposed approaches are very effective.

Categories and Subject Descriptors

H.1.2 [**Models and Principles**]: User/Machine Systems—*Human information processing*; H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval; I.5.3 [**Pattern Recognition**]: Clustering-algorithms

General Terms

Design, Experimentation, Algorithms, Measurement

Keywords

social tagging, tag search, tag clustering

SWSM'09, November 6, 2009, Hong Kong, China.

1. INTRODUCTION

Social tagging systems, such as del.icio.us[1] [8, 11], Flickr[2] [16], rawsugar[3] [14], citeulike[4] [25], and LibraryThing[5] have recently emerged as an effective and convenient way for users to label and organize large collections of resources. Compared to traditional centrally organized systems, such as university library system, which are based on controlled vocabularies, social tagging systems have the advantages that users can annotate resources with free-form tags. This leverages users' collaborative creation and understanding to resources [21], and is a more appropriate way to describe the ever-changing and wide-ranging corpora on the Web [30].

Although having difference in types of tagging resources, such as documents, URLs, images and so on, social tagging systems share basic functions. They allow users to bookmark resources, and resources can be retrieved by using any of the tags used to describe the resource. Besides the personal usage, an element of social interaction is also introduced. Social tagging systems allow users to share their publicly-saved resources and tags for these resources. The main service is called discovery and discovery means finding useful resources through tags. For example, a user who is interested in things about *Obama* can first search for the tag *obama* and then click certain links in that category. Thus, social tagging systems not only provide users with an effective way to enhance resources management, but also enable them to share vast amounts of resources.

After investigating del.icio.us, the most popular social tagging system for tagging URLs, we summarize the services that users can use to explore resources. Services provided by other social tagging systems are similar. Four services users can use are listed below:

1. Choose tags from a list of popular tags provided by the social tagging system.
2. Choose tags from tag clouds. Tag clouds are visual presentations of a set of tags. Attributes of the text such as position and size are used to represent features such as popularity of the associated tags.
3. Search in social tagging systems using one or several tags according to personal demands.

[1]http://delicious.com/
[2]http://flickr.com/
[3]http://www.rawsugar.com//
[4]http://www.citeulike.org/
[5]http://www.librarything.com/

4. Choose tags from a list of tags that are related to a tag the user is currently investigating.

According to the above four approaches, it can be concluded that currently social tagging systems have provided effective services for users who do not have some specific demands but just want to wander in the social tagging system in order to find some interesting things. However, some difficulties will occur when searching for specific concept in tagging systems. These difficulties mainly happens in two scenarios.

The first scenario is that when a user types a specific tag query into social tagging systems, he or she may find that not only very few results are returned, but also qualities of returned results are not satisfactory. This limitation comes from the fact that people from different countries or different cities all over the world may use different vocabularies to tag the same concept. Without loss of generality, taking English as an example, observing people coming from English-as-first-language countries, such as UK and USA, people from English-as-second-language countries, such as China and Japan, there is a large variation in vocabularies [4]. For example, in US, "mother tongue" is used to represent the concept "one's native language", while "mother language" is more commonly used to represent the same concept in China. Even people from countries with the same native language may tag a concept with different vocabularies, owing to their different language habits. Taking vocabularies in American English and British English for example, "chips" is used in British English while "french fries" is used in American English. There are plenty of similar vocabulary differences that can be found. Now that one concept can be described by different vocabularies, which are expressed as tags here, it is natural that some tags are used by a group of people while other tags are used by another group, though these two sets of tags can represent same concept. Also it is very likely that some tags may be less popular than others in describing the same concept. If a user unluckily types a less popular tag to explore, there is a good chance that he or she gets bad results.

The second scenario is that after getting results of a tag query, the user may want to know more if current results do not represent his or her interests [10]. On one side, a user may want to *roll-up* the concept of a tag query, meaning climbing up a concept hierarchy of a tag query. On the other side, the user is likely to *drill-down* the concept of a tag query, meaning navigating from less detailed data to more detailed data which can be realized by stepping down a concept hierarchy of a tag query. For example, after investigating results returned by typing *car*, the user may want to *drill-down* to a brand of car, for example, *BMW*.

In this paper, we propose a novel three-phase approach to overcome the above limitations. Our three-phase approach is based on the insightful investigation and observation on the user generated tags in real-world social tagging systems. Our three-phase approach is listed as follows:

1. Generating semantically-related tags for each *tag query*. For a given tag query, we will generate semantically-related tags for it, by "semantically-related tags" we mean tags that can be *rolled-up* from the concept hierarchy of this tag query, tags that can be *drilled-down* from the concept hierarchy of this tag query, or tags on the same concept hierarchy level of this tag query.
2. Constructing clusters of tags representing the same concept among semantically-related tags.
3. Building hierarchical relations among clusters of tags that represent the same concept.

This three-phase approach is proposed based on three reasonable assumptions:

Assumption 1. If two tags are semantically-related to each other, they may be used to annotate the same or similar resources. The idea of this assumption is related to [22, 24, 27]. Turney in [27] used the PMI-IR algorithm to determine whether two words are synonyms. This method used returned results of search engine and was based on the assumption that two similar words have a tendency to co-occur, which means two words can often cover same documents. Shen et al. in [24] judged whether two search queries are related based on the assumption that if two queries are related to each other, they should share some of the same or similar clicked Web pages. Because tags can represent human being's judgments about resources as claimed in [17, 31], it is natural for us make this assumption. With this assumption, we can generate the set of semantically-related tags for a tag query.

Assumption 2. If two tags can represent the same concept, not only are they usually used to annotate the same or similar resources, but also when tagging the same resource, the co-occurrening tags of these two tags used by different users should be similar. Taking annotating *URL* for example, for a *URL* talking about a new model of *bmw*, one user may use *car*, *bmw*, and *model* to annotate it, while another user may use *automobile*, *bmw*, and *model* to describe this. We propose our method for constructing clusters of tags representing the same concept based on this assumption.

Assumption 3. Our third assumption for building hierarchical relationships among clusters of tags that represent same concept is that if we have two clusters of tags, denoted as $c1$ and $c2$ here, $c1$ subsumes $c2$ if most objects annotated by tags in $c2$ also have similar objects annotated by tags in $c1$, while only part of objects annotated by tags in $c1$ have similar objects annotated by tags in $c2$. This assumption is related to the idea *term subsumption* in [20]. *Term subsumption* means that for two terms $t1$ and $t2$, $t1$ subsumes $t2$ if the documents in which $t2$ occurs are a subset of the documents in which $t1$ occurs.

The outline of the paper is as follows. Section 2 discusses the related work. Section 3 presents the details of our three-phase approach. Section 4 first analyzes the real-world data set we use for this paper, then presents and discusses our experimental results. Section 5 concludes the entire paper.

2. RELATED WORK

The vital growth of social tagging systems, which is a form of social media systems [18, 19], has created many interesting and challenging problems to the research community. Our work is most related to the problem of finding related tags and building hierarchical relationships among tags in social tagging systems.

Finding related tags in social tagging systems has been a hot research topic recently. Begelman et al. in [3] applied the "Co-tags" relation to construct a relation graph of

tags, and then recursively run a partition algorithm to construct a tag cluster to find related tags. Heymann et al. in [12] proposed tag-based association rules method to do tag prediction, giving a deeper understanding into the relationships between tags. Li et al. in [17] proposed that patterns of frequent co-occurrences of user tags can be used to characterize and capture topics of user interests, because these co-occurrence tags are related and can be grouped to represent topics. These approaches can only find related tags that have co-occurrence property, while failing to find related tags which do not co-occur. Our approach is different because our work is based on the observation that if tags are semantically-related to one another, they are usually used to annotate the same or similar resources.

In addition, several tasks have been proposed to build hierarchical relationships among tags. Kome in [15] showed that hierarchical relationships can be implicitly found in tagging systems. Heyman et al. in [10] presented a simple algorithm to automatically convert tags associated to objects into a hierarchical taxonomy. Schwarzkopf et al. in [23] proposed a way of creating taxonomic structure from a set of association rules. However, all these tasks only build hierarchical relationships between single tags. Our work is different from these tasks because our work builds hierarchical relationships among clusters of tags representing the same concept.

3. OUR APPROACH

In this section, we will first define some formulas for our approach, then in subsequent subsections, we will present our three-phase approach for building hierarchical relationships among clusters of tags representing the same concept in detail.

3.1 Preliminaries

Our three-phase approach for building hierarchical relations among clusters of tags that represent the same concept is the following: (1) Generate semantically-related tags for each tag query; (2) Construct clusters of tags that represent the same concept among related tags; (3) Build hierarchical relationships among the clusters of tags that represent the same concept.

Without loss of generality, in our formula we will use a special type of Web resource *URL* to represent the *resource* in order to increase the clarity of the paper.

We first define the set of URLs $URLset$, the set of users $USERset$, the set of tags $TAGset$. The data sets are defined as following:

$$URLset = \{url_i\}_{i=1}^{I},$$

$$USERset = \{user_j\}_{j=1}^{J},$$

$$TAGset = \{tag_x\}_{x=1}^{n}.$$

Here I is the total number of unique URLs in the tagging system, J is total number of unique users in the tagging system, and n is the total number of unique tags in the tagging system. Then we construct a vector of tags from TAGset:

$$tagvector = [tag_1, tag_2, \ldots, tag_x, \ldots, tag_n]^T,\ tag_x \in TAGset,$$

where tags in tagvector are sorted alphabetically.

For each URL in data set, there may be several users annotating it with one tag or a combination of several tags. In order to represent this phenomenon, we introduce a definition *Post*.

Post One post means one user uses one tag or a combination of several tags to annotate one URL. A formal expression of *Post* is expressed as follows:

$$post_{ij} = (url_i, user_j, tagset_{ij}),$$

where $tagset_{ij}$ is a set of tags, which records tags used by $user_j$ to annotate url_i.

Based on the definition of *Post*, we further use

$$P_i = \{post_{ij} | \forall j\}$$

to express the *post* set of url_i. These formulas will be used throughout the subsequent discussion.

3.2 Generating Semantically Related Tags

It has been argued that if two tags are semantically-related to each other, they may be used to annotate the same or similar URLs. So we can generate the set of related tags for a tag query. By "semantically-related tags" we mean tags that can be *rolled-up* or *drilled-down* from the concept hierarchy of a tag query or tags that are on the same concept hierarchy level of this tag query. Here we introduce a novel measure method *Coverage Rate* to measure whether two tags are semantically related.

Coverage Rate Given two tags tag_i and tag_j, the *Coverage Rate* of tag_i and tag_j measures how much resources covered by tag_i can also covered by tag_j. In other words, *coverage rate* measures how much resources annotated by tag_i have the same or similar URLs annotated by tag_j. Formal definition for *Coverage Rate* $CR(tag_i, tag_j)$ is defined by Eq. (1):

$$CR(tag_i, tag_j) = \frac{\sum_{n=1\ldots N} \max_{m=1\ldots M} sim(url_{in}, url_{jm})}{N}. \quad (1)$$

In Eq. (1), $url_{in} \in U(tag_i)$ and $url_{jm} \in U(tag_j)$, and $U(tag_i)$ means the set of URLs annotated by tag tag_i, and $U(tag_j)$ means the set of URLs annotated by tag tag_j. N is the total number of URLs annotated by tag_i, and M is the total number of URLs annotated by tag_j. $sim(url_{in}, url_{jm})$ measures the similarity between url_{in} and url_{jm}. Before discussing similarity measure method, we first need to develop an effective and efficient way to represent a URL. Different from traditional tf-idf representation, here we utilize millions of tags in social tagging systems to develop a new way to express a URL which we call it *Social Expression* of a URL.

Social Expression Social Expression utilizes users' understanding and description of URLs in the form of tags annotated with URLs, which leverages millions of tags in social tagging systems and is an effective and efficient way to describe URLs and there is even no need to analyze original Web contents of URLs. Formal expression of Social Expression of one URL $SE(url_i)$ is defined in Eq. (2):

$$SE(url_i) \in R^n, \quad (2)$$

where the x-th element, $SE(url_i)_{[x]}$, is

$$SE(url_i)_{[x]} = \begin{cases} 1 & \text{if } tag_x \in tagset_{url_i}, \\ 0 & \text{otherwise,} \end{cases}$$

and tag_x is the x-th element in tagvector with

$$tagset_{url_i} = \{tag_y | \ tag_y \ is \ used \ by \ at \ least \ one \ user \ to \ annotate \ url_i \ once.\}$$

In Eq. (2), elements in $SE(url_i)$ reflect which tags have been used by some user to annotate url_i. Because tags are a reflection of users' understanding and description of this URL. The fact that, tags are appropriate to represent human being's judgements about Web content has been argued by [17, 26].

Based on the definition of *Social Expression*, we can use some similarity measure method to calculate the similarity between two URLs, generally cosine similarity is used. Then, based on the understanding that semantically-related tags for a tag query are tags that can be *rolled-up* from the concept hierarchy of this tag query, tags that can be *drilled-down* from the concept hierarchy of this tag query or tags that are on the same concept hierarchical level of this tag query, we can come up with a method to judge whether a tag_j is semantically-related to tag_i. The method is define in Eq. (3):

$$CR(tag_i, tag_j) > \theta_1 \ and \ CR(tag_j, tag_i) > \theta_1. \quad (3)$$

Requirement of Eq. (3) not only greatly avoid the effect of noise in social tagging systems, but also it has considered concept of tag_j can be *rolled-up* from the concept hierarchy of tag_i, or concept of tag_j can be *drilled-down* from the concept hierarchy of tag_i, or concept of tag_j is on the same hierarchy level of concept hierarchy of tag_i. It implies that two tags are semantically-related in one social tagging system because there are a certain amount of URLs annotated by one tag also have similar URLs annotated by another tag.

3.3 Constructing Clusters of Tags Representing Same Concept

Based on assumption 2, if two tags can represent one same concept, not only are they usually used to annotate the same or similar resources, but also when tagging a same resource, the co-occurring tags of these two tags used by different users should be similar. In the following section, we will show details of our approach for constructing clusters of tags that represent same concept.

In social tagging systems, when tag_k (the k-th tag in tagvector) is used by $user_j$ to describe url_i, it is very likely that several other tags are also used by the $user_j$ to annotate url_i. We define these co-occurring tags of tag_k in one *Post* $post_{ij}$ as tag_k's context in this post $post_{ij}$. Here we give a formal definition of context of tag_k in $post_{ij}$ in Eq. (4):

$$ct_{kij} \in R^n, \quad (4)$$

where the x-th element $ct_{kij[x]}$ is defined as

$$ct_{kij[x]} = \begin{cases} 1 & \text{if } tag_x \in tagset_{ij}, k \neq x, \\ 0 & \text{otherwise,} \end{cases}$$

and tag_x is the x-th element in tagvector with

$$tagset_{ij} = \{tag_y | tag_y \ is \ used \ by \ user_j \ to \ annotate \ url_i.\}$$

For two tags tag_k and tag_m, if they can represent the same concept and they are used by two different users $user_n$ and $user_o$ to annotate one URL url_i, it is very likely that ct_{kin}, which is the context of tag_k in $post_{in}$ is very similar to ct_{mio}, which is the context of tag_m in $post_{io}$. Based on this observation, we can define the context-based similarity, $simct(tag_{ki}, tag_{mj})$, between tag_{ki} and tag_{mj} as follows:

$$simct(tag_{ki}, tag_{mj}) = \begin{cases} \frac{\sum_{n=1...N} \max_{o=1...O} sim(ct_{kin}, ct_{mjo})}{N} & i = j, n \neq o \\ 0 & i \neq j \end{cases}. \quad (5)$$

In Eq. (5), tag_{ki} means tag_k is annotated by at least one user to URL url_i, tag_{mj} means tag_m is annotated by at least one user to URL url_j. Meanings of ct_{kin} and ct_{mjo} have been defined in Eq. (4). N is number of users who use tag_k to annotate url_i and O is number of users who use tag_m to annotate url_j. Equation (5) measures similarity between two tags for a URL given two tags' separate context. In Eq. (5), $sim(ct_{kin}, ct_{mjo})$ measures similarity between two context, according to Eq. (4), ct_{kin} is context of tag_k in $post_{in}$ and ct_{mjo} is context of tag_m in $post_{jo}$. There are a number of similarity measures for this, generally cosine similarity will be considered. Now, based on Eq. (5), we can define the context-based similarity measure $context\_sim(tag_k, tag_m)$ between two tags tag_k and tag_m in Eq. (6).

$$context\_sim(tag_k, tag_m) = \frac{\sum_{i=1...I} \sum_{j=1...J} simct(tag_{ki}, tag_{mj}) + \sum_{j=1...J} \sum_{i=1...I} simct(tag_{mj}, tag_{ki})}{2NC}. \quad (6)$$

In Eq. (6), NC means number of times both tag_k and tag_m are used to annotate a same *URL*.

Also, we can know that if two tags can represent the same concept, they may be used to annotate some same URLs, so we also consider this factor. Now we can define the similarity measure $sim(tag_k, tag_m)$ between tags tag_k and tag_m in Eq. (7).

$$sim(tag_k, tag_m) = a \times \frac{\frac{NC}{N} + \frac{NC}{O}}{2} + (1 - a) \times context\_sim(tag_k, tag_m). \quad (7)$$

In Eq. (7), NC means number of times both tag_k and tag_m are used to annotate a same *URL*. N is number of URLs that has been annotated with tag_k, O is number of URLs that has been annotated with tag_m. The context-based similarity measure $context\_sim(tag_k, tag_m)$ has been defined in Eq. (6). So Eq. (7) can give an approximation to measure similarity between tag_k and tag_m. After proposing similarity measure function between two tags, we can compute the similarity score between pairs of tags and construct a similarity matrix, then we use a state-of-the-art clustering technique, correlation clustering to automatically cluster the tags. Correlation clustering contains the advantage of determining the optimal number of clusters k without specifying beforehand [2], it has been applied to some applications successfully [6, 7].

3.4 Building Hierarchical Relations Among Clusters of Tags

Following our third assumption, We can determine hierarchical relationships among clusters of tags. For two clusters of tags, denoted as $c1$ and $c2$ here, $c1$ subsumes $c2$ if most resources annotated by tags in $c2$ also have similar resources annotated by tags in $c1$, while only part of resources annotated by tags in $c1$ have similar resources annotated by tags in $c2$. In the following section, we will formulate this idea. Because we have defined the concept of *Coverage Rate*, we can use it to generate the rule determining hierarchical relationships among clusters of tags. First we will formulate the *Coverage Rate* between two clusters in Eq. (8).

$$CR(c_i, c_k) = \frac{\sum_{j=1...J} \sum_{m=1...M} CR(t_{ij}, t_{km})}{J \times M}, t_{ij} \in c_i, t_{km} \in c_k. \quad (8)$$

Based on Eq. (8), we can get the rule for determining hierarchical relations between two clusters.

$$cluster c_i \text{ subsumes cluster } c_k \text{ if } CR(c_k, c_i) - CR(c_i, c_k) > \theta_2. \quad (9)$$

$$cluster c_k \text{ subsumes cluster } c_i \text{ if } CR(c_i, c_k) - CR(c_k, c_i) > \theta_2. \quad (10)$$

$$\text{Otherwise } c_i \text{ and } c_k \text{ is on the same level.} \quad (11)$$

4. EXPERIMENTATION

4.1 Data Set

4.1.1 Data Collection

The data used for this paper was obtained by systematically crawling the del.icio.us portals during December of 2006. del.icio.us is a very popular social tagging system. In del.icio.us, users can bookmark their interested URLs, then add one tag or several tags to describe the bookmark according to their personal understanding. If the bookmark is publicly-saved, tags added to it also can be used for search by other users. In addition, users can add their own tags to the bookmark pointing to the same URLs independently.

In the crawling activity, each bookmark we crawl consists of URL, user, and tags annotated to the URL by this user. For all the users contained in the data set, almost complete information about their postings was collected. The basic statistics of the crawled data are summarized in Table 1.

Table 1: Statistics of the crawled data sets

users	tags	resources
532,924	2,473,657	17,262,475

4.1.2 Distribution Analysis

After collecting the data, we analyze distribution of the data. We analyze our data from three aspects, users' participation, URLs' annotation, and tags' usage frequencies.

Users' Participation. As in the majority of online communities, users' participation is unevenly distributed. Figure 1(a) shows the distribution of users' participation.

Figure 1(a) shows the frequency distribution of the users' annotations in the log-log scale. Here one annotation means an activity that one user uses a tag or a combination of several tags to annotate one URL. We can see that the majority of users finish a small number of annotations, while a handful of users accomplish the most annotations. The distribution matches the standard skewed distributions of other activities in online spaces such as Witkey sites [28], wikis [13], question answer forums [1], and search activities [5, 9].

URLs' annotation. Figure 1(b) shows the distribution of URLs' annotation activity in a log-log scale. Here one annotation means a URL is annotated by a user with one tag or a combination of several tags.

In Fig. 1(b), we can see that the distribution of points follows the power law distribution, which means that the majority of URLs are less annotated, while a small portion of URLs are annotated more often.

Tags' usage frequency. Figure 1(c) shows the distribution of tags' usage frequency. In the figure, the meaning of one annotation is that a tag is used by a user to annotate a URL.

4.2 Generating Semantically Related Tags

The goal of the step of generating semantically-related tags is to generate the set of related tags for a tag query. In the paper, semantically-related tags mean tags that can be *rolled-up* from the concept hierarchy of this tag query, tags that can be *drilled-down* from the concept hierarchy of this tag query, and tags that are on the same level of the concept of this tag query. To study the effectiveness our method for generating semantically-related tags, we randomly sampled 15 *seed tags* from the data set, these 15 *seed tags* come from concepts of several fields, and without loss of generality, all these 15 words are English words. Several semantically related tags for each *seed tag* are randomly sampled, resulting in totally 64 tags. These 64 have potential hierarchical relationships among them. Three human experts are invited to label whether a tag is semantically-related to another tag based on their personal understanding. A ground truth labeling result on whether a tag is semantically-related to another is developed based on a majority voting strategy that applied to three human experts' labeling results. Table 2 shows some examples of semantically-related tags for some *seed tags*.

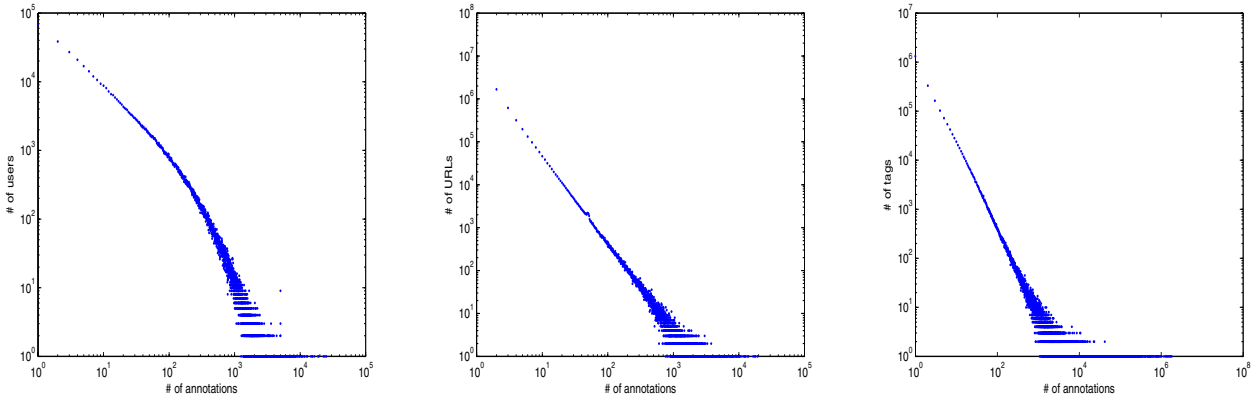
Table 2: Examples of semantically-related tags for tag queries

Tag query	Sample semantically-related tags
nikon	panasonic, camera
automobile	audi, bmw, car
windows	pc, tool, os, unix,computer

For all the sampled 64 tags, we use our approach to generate semantically-related tags for each tag. Figure 2(a) shows the accuracy of our method for generating semantically-related tags when we change the threshold. We use Eq. (12) to calculate the accuracy:

$$accuracy = \frac{N_c}{N_t}, \quad (12)$$

where N_c is total number of correct generated semantically-related tags for all tag queries. N_t is total number of generated semantically-related tags for all tag queries. Figure



(a) The distribution of users' participation.

(b) The distribution of URLs' annotation.

(c) The distribution of tags' usage frequency.

Figure 1: Distribution analysis of dataset.

2(b) shows the average number of generated semantically-related tags for each tag when changing the threshold. We can see that our methods to generate semantically-related tags achieves promising performance.

4.3 Constructing Clusters of Tags Representing Same Concept

After testing our method of generating semantically-related tags for tag queries, we test our approach to construct clusters of tags representing same concept among each group of semantically-related tags generated by our approach. Here we select three sample groups of semantically-related tags generated by our approach that are from three different domains, to demonstrate our method's effectiveness in constructing clusters of tags representing same concept. These three groups of tags are listed in Table 3. The first tag in each group is the tag query, which means that other tags in its group is its semantically-related tags generated by our method.

Table 3: Three groups of semantically-related tags generated by our approach

Domain	Groups of semantically-related tags
Autos	**automobile**, audi, bmw, car
Shopping	**fashion**, shop, store, style, clothes
Computer	**windows**, pc, tool, os, unix,computer

According to Eq. (7), we can know that the similarity measure between two tags tag_k and tag_m are consisted of two parts, the first part reflects percentage of number of URLs that are annotated by two tags, the second part reflect the similarity of tag context. In our experiment, we found an interesting observation that if two tags are both on a high concept level and can represent the same concept, which means that two tags are relatively abstract concepts, such as *automobile* and *car* comparing to *bmw* and *audi*, they will have a relatively high percentage of URLs that have been annotated with both tags, in other words, the similarity score in the first part is relatively big; if two tags are both on a low concept level and can represent the same concept, such as *bmw* and *audi* comparing to *automobile* and *car*, they will have a relatively high context similarity, which is the second part of Eq. (7); if two tags are from different concept level, they will have ordinary similarity score in both parts. Based on this meaningful observation, we apply a heuristic way to dynamically change the value of a to adjust weights of two parts in Eq. (7). Figure 3 demonstrates results of constructing clusters of tags representing same concept for three sample groups of semantically-related tags.

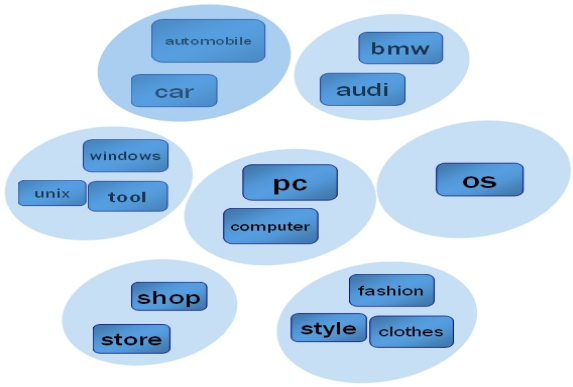
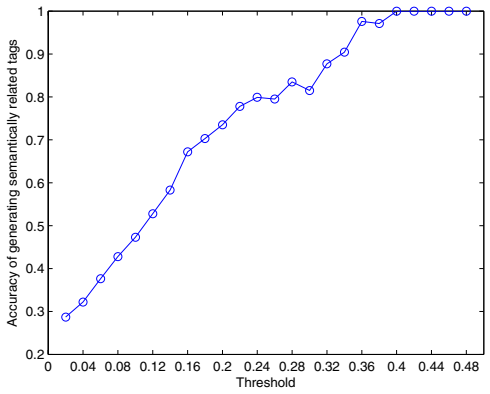


Figure 3: Constructing clusters of tags representing same concept among sample groups

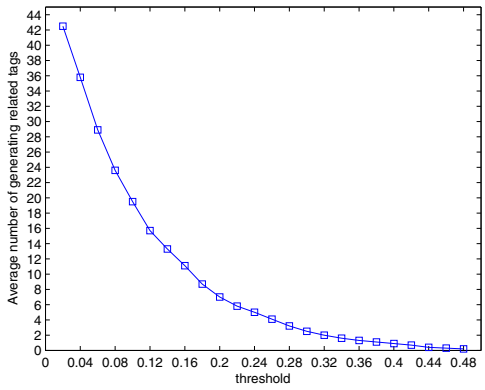
From the result we can see that our method of constructing clusters of tags representing same concept is very effective.

4.4 Building Hierarchical Relations Among Clusters of Tags

To test our method's effectiveness in building hierarchical relations among clusters of tags, we use previous experiment's results as this experiment's input, and check whether hierarchical relations in results can reflect real world's concept hierarchy. We choose the threshold in Eq. (9) and Eq. (10) as 0.03. Figure 4 demonstrates results of building hierarchical relations among clusters of tags.



(a) The accuracy of generating semantically-related tags.



(b) Average number of generated semantically-related tags.

Figure 2: Results of generating semantically related tags.

Figure 4: Building hierarchical relations among clusters of tags

From the results of building hierarchical relations among clusters of tags, we can find only one result that does not reflect the real world's concept hierarchy. It is cluster of tags {windows, tool, unix} subsumes cluster of tag {os}, which is different from the view of real world's concept hierarchy. After checking the data set, we find that the tag *windows* is not only used to annotate a URL which mainly discusses this type of operating system, but also it is often used to annotate URLs that discuss software that runs on *windows* platform, then in this situation a tag that represents a higher concept such as *os* won't be used. The phenomenon is similar for tag *unix*. However, this is a rare failure in the dataset, it only happens when a tag is used to annotate URLs in which this tag does not reflect the key point of the URL but rather is used to help remember other tags that reflect the key point of the URL.

Remark. It may be argued that the effectiveness of our methods may be affected by the problem of tag ambiguity. However, recently, Yeung et al. in [29] proposes a k-Nearest-Neighbor Method to tackle the problem of tag ambiguity in social tagging systems, an ambiguous tag's accurate meaning can be inferred with the help of its co-occurred tags. Our three-phase approach is extensible to combine its method to solve the problem of tag ambiguity. For example, a tag "apple" is transformed to "apple-fruit" if it occurs with "pie", while transformed to "apple-company" if it occurs with "iPhone".

5. CONCLUSION AND FUTURE WORK

In this paper, we have proposed a three-phase approach, which is generating semantically-related tags, constructing clusters of tags representing same concept, and building hierarchical relations among clusters of tags, to explore tagging wisdom. We put forward the concept *Coverage Rate* which is shown to be useful in measuring whether two tags are semantically-related. We also put forward the concept of a tag's *context* to reflect an activity that a user uses this tag and several other tags to annotate a URL, and tags' *context* is utilized to help construct clusters of tags representing same concept. The experimental results employing the real world data set help validate the effectiveness of our methods.

We plan to extend our methods to tackle the problem of tag ambiguity in our future work. We also hope to apply our methods to more applications including tag query suggestion and tag query expansion.

Acknowledgment

The authors would like to thank the anonymous reviewers providing helpful comments. We thank Mr. Haiqin Yang for many valuable discussions on this topic. The work described in this paper is supported by grants from the Research Grant Council of the Hong Kong Special Administrative Region, China (Project No.: CUHK 4128/08E and CUHK 4158/08E) and Microsoft (Project No.: 6902498). This work is also affiliated with the Microsoft-CUHK Joint Laboratory for Human-centric Computing and Interface Technologies.

6. REFERENCES

[1] L. A. Adamic, J. Zhang, E. Bakshy, and M. S. Ackerman. Knowledge sharing and yahoo answers: everyone knows something. In *Proceedings of the 17th*

International Conference on World Wide Web, pages 665–674, 2008.

[2] N. Bansal, A. Blum, and S. Chawla. Correlation clustering. *Machine Learning*, 56(1-3):89–113, 2004.

[3] G. Begelman, P. Keller, and F. Smadja. Automated Tag Clustering: Improving search and exploration in the tag space. In *Proceedings of the Collaborative Web Tagging Workshop at WWW 2006*, 2006.

[4] D. Crystal. The future of Englishes. *English Today*, 15(02):10–20, 2008.

[5] H. Deng, I. King, and M. R. Lyu. Entropy-biased models for query representation on the click graph. In *Proceedings of the 32nd Annual ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 339–346, 2009.

[6] T. Finley and T. Joachims. Supervised clustering with support vector machines. In *Proceedings of the 22nd International Conference on Machine Learning*, pages 217–224, 2005.

[7] J. V. Gael and X. Zhu. Correlation clustering for crosslingual link detection. In *Proceedings of the 20th International Joint Conference on Artificial Intelligence*, pages 1744–1749, 2007.

[8] S. A. Golder and B. A. Huberman. The structure of collaborative tagging systems. *CoRR*, abs/cs/0508082, 2005.

[9] Q. He, D. Jiang, Z. Liao, S. C. Hoi, K. Chang, E.-P. Lim, and H. Li. Web query recommendation via sequential query prediction. In *Proceedings of 25th International Conference on Data Engineering*, 2009.

[10] P. Heymann and H. Garcia-Molina. Collaborative creation of communal hierarchical taxonomies in social tagging systems. Technical report, Technical Report 2006-10, Stanford University, April 2006.

[11] 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.

[12] P. Heymann, D. Ramage, and H. Garcia-Molina. Social tag prediction. In *Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 531–538, 2008.

[13] T. Holloway, M. Bozicevic, and K. Börner. Analyzing and visualizing the semantic coverage of wikipedia and its authors. *Complexity*, 12(3):30–40, 2007.

[14] S. Kelkar, A. John, and D. Seligmann. An activity-based perspective of collaborative tagging. In *International Conference on Weblogs and Social Media*, 2007.

[15] S. Kome. Hierarchical Subject Relationships in Folksonomies. 2005.

[16] J. Leskovec, L. Backstrom, R. Kumar, and A. Tomkins. Microscopic evolution of social networks. In *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 462–470, 2008.

[17] X. Li, L. Guo, and Y. E. Zhao. Tag-based social interest discovery. In *Proceedings of the 17th International Conference on World Wide Web*, pages 675–684, 2008.

[18] H. Ma, I. King, and M. R. Lyu. Effective missing data prediction for collaborative filtering. In *Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 39–46, 2007.

[19] H. Ma, I. King, and M. R. Lyu. Learning to recommend with social trust ensemble. In *Proceedings of the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 203–210, 2009.

[20] M. Sanderson. Deriving concept hierarchies from text. In *Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 206–213, 1999.

[21] R. Schenkel, T. Crecelius, M. Kacimi, S. Michel, T. Neumann, J. Parreira, and G. Weikum. Efficient top-k querying over social-tagging networks. In *Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 523–530, 2008.

[22] P. Schmitz. Inducing ontology from flickr tags. In *Proceedings of Collaborative Web Tagging Workshop at WWW 2006*, 2006.

[23] E. Schwarzkopf, D. Heckmann, D. Dengler, and A. Kröner. Mining the structure of tag spaces for user modeling. In *Proceedings of the Workshop on Data Mining for User Modeling at the 11th International Conference on User Modeling*, pages 63–75, 2007.

[24] D. Shen, M. Qin, W. Chen, Q. Yang, and Z. Chen. Mining web query hierarchies from clickthrough data. In *Proceedings of the 22nd AAAI Conference on Artificial Intelligence*, pages 341–346, 2007.

[25] Y. Song, Z. Zhuang, H. Li, Q. Zhao, J. Li, W. Lee, and C. Giles. Real-time automatic tag recommendation. In *Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 515–522, 2008.

[26] F. M. Suchanek, M. Vojnovic, and D. Gunawardena. Social tags: Meanings and suggestions. In *ACM 17th Conference on Information and Knowledge Management*, pages 223–232.

[27] P. D. Turney. Mining the web for synonyms: Pmi-ir versus lsa on toefl. *Lecture Notes in Computer Science*, pages 491–502, 2001.

[28] J. Yang, L. Adamic, and M. Ackerman. Competing to share expertise: the taskcn knowledge sharing community. In *Proceedings of International Conference on Weblogs and Social Media*, 2008.

[29] C. M. A. Yeung, N. Gibbins, and N. Shadbolt. A k-nearest-neighbour method for classifying web search results with data in folksonomies. pages 70–76, 2008.

[30] D. Zhou, J. Bian, S. Zheng, H. Zha, and C. L. Giles. Exploring social annotations for information retrieval. In *Proceedings of the 17th International Conference on World Wide Web*, pages 715–724, 2008.

[31] T. C. Zhou, H. Ma, I. King, and M. R. Lyu. Tagrec: Leveraging tagging wisdom for recommendation. In *Proceedings of IEEE International Symposium on Social Intelligence and Networking*, 2009.