Exploring Social Annotations for the Semantic Web

Xian Wu*
Shanghai JiaoTong University
Shanghai, 200030, China
wuxian@apex.sjtu.edu.cn

Lei Zhang
IBM China Research Lab
Beijing, 100094, China
lzhangl@cn.ibm.com

Yong Yu
Shanghai JiaoTong University
Shanghai, 200030, China
yyu@apex.sjtu.edu.cn

ABSTRACT

In order to obtain a machine understandable semantics for web resources, research on the Semantic Web tries to annotate web resources with concepts and relations from explicitly defined formal ontologies. This kind of formal annotation is usually done manually or semi-automatically. In this paper, we explore a complement approach that focuses on the "social annotations of the web" which are annotations manually made by normal web users without a predefined formal ontology. Compared to the formal annotations, although social annotations are coarse-grained, informal and vague, they are also more accessible to more people and better reflect the web resources' meaning from the users' point of views during their actual usage of the web resources. Using a social bookmark service as an example, we show how emergent semantics [2] can be statistically derived from the social annotations. Furthermore, we apply the derived emergent semantics to discover and search shared web bookmarks. The initial evaluation on our implementation shows that our method can effectively discover semantically related web bookmarks that current social bookmark service can not discover easily.

Categories and Subject Descriptors

H.4.m [**Information Systems**]: Miscellaneous

General Terms

Alogrithms, Experimentation

Keywords

semantic web, social annotation, emergent semantics, social bookmarks

1. INTRODUCTION

Semantic Web is a vision that web resources are made not only for humans to read but also for machines to understand and automatically process [3]. This requires that web resources be annotated with machine understandable metadata. Currently, the primary approach to achieve this is to firstly define an ontology and then use the ontology to add semantic markups for web resources. These semantic markups are written in standard languages such as RDF [20] and OWL [23] and the semantics is provided by the ontology that is shared among different web agents and applications. Usually, the semantic annotations are made manually using a toolkit such as Protege or CREAM [26, 31] or semi-automatically through user interaction with a disambiguation algorithm [18, 4, 5, 6]. There are also some work on automatic annotation with minimum human efforts. They either extract metadata from the web site's underlying databases [12] or analyze text content within the web pages using learning algorithms [7] and/or NLP techniques [8]. Most of these methods uses a pre-defined ontology as the semantic model for the annotations. The manual and semi-automatic methods usually requires the user be familiar with the concept of ontologies and taxonomies. Although these approaches have been successfully used in applications like bioinformatics (e.g. [22]) and knowledge management (e.g. [18]), they also have some disadvantages. Firstly, establishing an ontology as a semantic backbone for a large number of distributed web resources is not easy. Different people/applications may have different views on what exists in these web resources and this leads to the difficulty of the establishment of an commitment to a common- ontology. Secondly, even if the consensus of a common ontology can be achieved, it may not be able to catch the fast pace of change of the targeted web resources or the change of user vocabularies in their applications. Thirdly, using ontologies to do manual annotation requires the annotator have some skill in ontology engineering which is a quite high requirment for normal web users.

In this paper, we explore a complement approach of semantic annotations that focuses on the "social annotations" of the web. In the recent years, web blogs and social bookmarks services are becoming more and more popular on the web. A web blog service usually allows the user to categorize the blog posts under different category names chosen by the user. Social bookmark services (e.g. del.icio.us[1]) enable users to not only share their web bookmarks but also assign "tags" to these bookmarks. These category names and tags are freely chosen by the user without any a-priori dictionary, taxonomy, or ontology to conform to. Thus, they can be any strings that the user deems appropriate for the web resource. We see them as the "social annotations" of the web. We use the word "social" to emphasize that these annotations are made by a large number of normal web users

*Part of Xian Wu's work of this paper was conducted in IBM China Research Lab.

WWW 2006, May 23–26, 2006, Edinburgh, Scotland.
ACM 1-59593-323-9/06/0005.

[1]http://del.icio.us

with implicit social interactions on the open web without a pre-defined formal ontology. Social annotations remove the high barrier to entry because web users can annotate web resources easily and freely without using or even knowing taxonomies or ontologies. It directly reflects the dynamics of the vocabularies of the users and thus evolves with the users. It also decomposes the burden of annotating the entire web to the annotating of interested web resources by each individual web users.

Apparently, without a shared taxonomy or ontology, social annotations suffer the usual problem of ambiguity of semantics. The same annotation may mean different things for different people and two seemingly different annotations may bear the same meaning. Without a clear semantics, these social annotations won't be of much use for web agents and applications on the Semantic Web. In this paper, using a social bookmark service as an example, we propose to use a probabilistic generative model to model the user's annotation behavior and to automatically derive the emergent semantics [2] of the tags. Synonymous tags are grouped together and highly ambiguous tags are identified and separated. The relationship with the formal annotations is also discussed. Furthermore, we apply the derived emergent semantics to discover and search shared web bookmarks and describe the implementation and evaluation of this application.

2. SOCIAL BOOKMARKS AND SOCIAL ANNOTATIONS

The idea of a social approach to the semantic annotation is enlightened and enabled by the now widely popular social bookmarks services on the web. These services provide easy-to-use user interfaces for web users to annotate and categorize web resources, and furthermore, enable them to share the annotations and categories on the web. For example, the Delicious (http://del.icio.us) service

> "allows you to easily add sites you like to your personal collection of links, to categorize those sites with keywords, and to share your collection not only between your own browsers and machines, but also with others" – [29]

There are many bookmarks manager tools available [17, 11]. What's special about the social bookmarks services like Delicious is their use of keywords called "tags" as a fundamental construct for users to annotate and categorize web resources. These tags are freely chosen by the user without a pre-defined taxonomy or ontology. Some example tags are "blog", "mp3", "photography", "todo" etc. The tags page of the Delicious web site (http://del.icio.us/tags/) lists most popular tags among the users and their relative frequency of use. These user-created categories using unlimited tags and vocabularies was coined a name "folksonomy" by Thomas Vander Wal in a discussion on an information architecture mailing list [32]. The name is a combination of "folk" and "taxonomy".

As pointed out in [21], folksonomy is a kind of user creation of metadata which is very different from the professional creation of metadata (e.g. created by librarians) and author creation of metadata (e.g. created by a web page author). Without a tight control on the tags to use and some expertise in taxonomy building, the system soon runs into problems caused by ambiguity and synonymy. [21] cited some examples of ambiguous tags and synonymous tags in Delicious. For example, the tag "ANT" is used by many users to annotate web resources about Apache Ant, a building tool for Java. One user, however, uses it to tag web resources about "Actor Network Theory". Synonymous tags, like "mac" and "macintosh", "blog" and "weblog" are also widely used.

Despite the seemingly chaos of unrestricted use of tags, social bookmarks services still attract a lot of web users and provide a viable and effective mechanism for them to organize web resources. [21] contributes the success to the following reasons.

- Low barriers to entry
- Feedback and Asymmetric Communication
- Individual and Community Aspects

Unlike the professional creation of metadata or the formal approach of the semantic annotation, folksonomy does not need sophisticated knowledge about taxonomy or ontology to do annotation and categorization. This significantly lowers the barrier to entry. In addition, because these annotations are shared among all users in a social bookmark service, there is an immediate feedback when a user tags a web resource. The user can immediately see other web resources annotated by other users using the same tag. These web resources may not be what the user expected. In that case, the user can adapt to the group norm, keep your tag in a bid to influence the group norm, or both [34]. Thus, the users of folksonomy are negotiating the meaning of the terms in an implicit asymmetric communication. This local negotiation, from the emergent semantics perspective, is the basis that leads to the incremental establishment of a common global semantic model. [24] made a good analogy with the "desire lines". Desire lines are the foot-worn paths that sometimes appear in a landscape over time. The emergent semantics is like the desire lines. It emerges from the actual use of the tags and web resources and directly reflects the user's vocabulary and can be used back immediately to serve the users that created them. In the rest of the paper, we quantitatively analyze social annotations in the social bookmarks data and show that emergent semantics indeed can be inferred statistically from it.

3. DERIVING EMERGENT SEMANTICS

In social bookmarks services, an annotation typically consists of at least four parts: the link to the resource (e.g. a web page), one or more tags, the user who makes the annotation and the time the annotation is made. We thus abstract the social annotation data as a set of quadruple

$$(user, resource, tag, time)$$

which means that a user annotates a resource with a specific tag at a specific time. In this paper, we focus on who annotates what resource with what tag and do not care much about the time the annotation is made. What interests us is thus the co-occurrence of users, resources and tags. Let's denote the set $U = \{u_1, u_2, \ldots, u_K\}$, $R = \{r_1, r_2, \ldots, r_M\}$, $T = \{t_1, t_2, \ldots, t_N\}$ to be the set of K users, M web resources and N tags in the collected social annotation data

respectively. Omitting the time information, we can translate each quadruple to a triple of $(user, resource, tag)$. As mentioned in Section 2, the social annotations are made by different users without a common dictionary. Hence, the problem of how to group synonymous tags, how to distinguish the semantics of an ambiguous tag becomes salient for sematic search. In this section, we use a probabilistic generative model to obtain the emergent semantics hidden behind the co-occurrences of web resources, tags and users, and implement semantic search based on the emergent semantics.

3.1 Exploiting Social Annotations

After analyzing a large amount of social annotations, we found that tags are usually semantically related to each other if they are used to tag the same or related resources for many times. Users may have similar interests if their annotations share many semantically related tags. Resources are usually semantically related if they are tagged by many users with similar interests. This domino effect on semantic relatedness also can be observed from other perspectives. For example, tags are semantically related if they are heavily used by users with similar interests. Related resources are usually tagged many times by semantically related tags and finally users may have similar interests if they share many resources in their annotations. This chain of semantic relatedness is embodied in the different frequencies of co-occurrences among users, resources and tags in the social annotations. These frequencies of co-occurrences give expression to the implicit semantics embedded in them.

Inspired by research on Latent Semantic Index [30], we try to make statistical studies on the co-occurrence numbers. We represent the semantics of an entity (a web resource, a tag or a user) as a multi-dimensional vector where each dimension represents a category of knowledge. Every entity can be mapped to a multi-dimensional vector, whose component on each dimension measures the relativity between the entity and the corresponding category of knowledge. If one entity relates to a special category of knowledge, the corresponding dimension of its vector has a high score. For example, in Del.icio.us, the tag 'xp' is used to tag web pages about both 'Extreme Programming' and 'Window XP'. Its vector thus should have high score on dimensions of 'software' and 'programming'. This actually is what we get in our experiments in Section 3.2. As in each annotation, the user, tag and resource co-occur in the same semantic context. The total knowledge of users, tags and resources are the same for them. Hence we can represent the three entities in the same multi-dimensional vector space, which we call the conceptual space. As illustrated in Fig.1, we can map users, web resources and tags to vectors in this conceptual space. For an ambiguous tag, it may have several notable components on different dimensions while a definite tag should only has one prominent component. In short, we can use the vectors in this conceptual space to represent the semantics of entities. Conceptual space is not a new idea. It also appears in many literatures studying e.g. the meaning of words [33] and texts [30].

Our job next is to determine the number of dimensions and acquire the vector values of entities from their co-occurrences. There are research on the statistical analysis of co-occurrences of objects in unsupervised learning. These approaches aim to first develop parametric models, and then estimate parameters by maximizing log-likelihood on the existing data set. The acquired parameter values can then be used to predict probability of future co-occurrences. Mixture models [14] and clustering models based on deterministic annealing algorithm [27] are of this kind approaches which have been used in many fields such as Information Retrieval [13] and Computational Linguistics [9]. We applied Separable Mixture Model [14](one kind of mixture models mentioned above) to the co-occurrence of tags and resources without users before in a separate paper [36]. In this paper, we extend the bigram Separable Mixture Model to a tripartite probabilistic model to obtain the emergent semantics contained in the social annotations data.

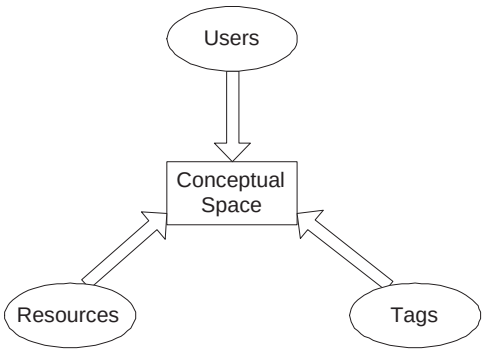

Figure 1: Mapping entities in folksonmies to conceptual space

We assume that the conceptual space is a D dimensional vector space, each dimension represent a special category of knowledge included in social annotation data. The generation of existing social annotation data can be modeled by the following probabilistic process:

1. Choose a dimension d_α to represent a category of knowledge according to the probability $p(d_\alpha), \alpha \in [1, D]$.
2. Measure the relativity between the interest of user u_i and the chosen dimension with the conditional probability $p(u_i|d_\alpha)$.
3. Measure the relativity between the semantics of a resource r_j and the chosen dimension with conditional probability $p(r_j|d_\alpha)$.
4. Measure the relativity between the semantics of a tag t_k and the chosen dimension according to the conditional probability $p(t_k|d_\alpha)$.

In the above model, the probability of the co-occurrence of u_i, r_j and t_k is thus:

$$p(u_i, r_j, t_k) = \sum_{\alpha=1}^{D} p(d_\alpha)p(u_i|d_\alpha)p(r_j|d_\alpha)p(t_k|d_\alpha) \quad (1)$$

The log-likelihood of the annotation data set is thus:

$$L = \sum_i \sum_j \sum_k n_{ijk} \log \sum_{\alpha=1}^{D} p(d_\alpha)p(u_i|d_\alpha)p(r_j|d_\alpha)p(t_k|d_\alpha) \quad (2)$$

where n_{ijk} denotes the co-occurrence times of u_i,r_j and t_k.

Probabilities in 2 can be estimated by maximizing the log-likelihood L using EM (Expectation-Maximum) method.

Suppose that the social annotations data set contains C triples. Let $u(c)$, $r(c)$, $t(c)$ denote the cth record in the data set containing the $u(c)$th user , the $r(c)$th resource and the $t(c)$th tag in respective set of users, resources, and tags. The $C*D$ matrix I is the indicator matrix of EM algorithm. $I_{c\alpha}$ denote the probability of assigning the cth record to dimension α.

E-step:

$$I_{c\alpha}^{(t)} = \frac{p(d_\alpha)^{(t)} p(u_{u(c)}|d_\alpha)^{(t)} p(t_{t(c)}|d_\alpha)^{(t)} p(r_{r(c)}|d_\alpha)^{(t)}}{\sum_{\alpha=1}^{D} p(d_\alpha)^{(t)} p(u_{u(c)}|d_\alpha)^{(t)} p(t_{t(c)}|d_\alpha)^{(t)} p(r_{r(c)}|d_\alpha)^{(t)}} \tag{3}$$

M-step:

$$p(d_\alpha)^{(t+1)} = \frac{\sum_{c=1}^{C} I_{c\alpha}^{(t)}}{C} \tag{4}$$

$$p(u_i|d_\alpha)^{(t+1)} = \frac{\sum_{c:u(c)=i} I_{c\alpha}^{(t)}}{\sum_{c=1}^{C} I_{c\alpha}^{(t)}} \tag{5}$$

$$p(r_j|d_\alpha)^{(t+1)} = \frac{\sum_{c:r(c)=j} I_{c\alpha}^{(t)}}{\sum_{c=1}^{C} I_{c\alpha}^{(t)}} \tag{6}$$

$$p(t_k|d_\alpha)^{(t+1)} = \frac{\sum_{c:t(c)=k} I_{c\alpha}^{(t)}}{\sum_{c=1}^{C} I_{c\alpha}^{(t)}} \tag{7}$$

Iterating E-step and M-step on the existing data set, the log-likelihood converges to a local maximum gradually, and we get the estimated values of $p(d)$, $p(u|d)$, $p(r|d)$ and $p(t|d)$. We can use these values to calculate the vectors of users, resources and tags in conceptual space using Bayes' theorem. For example, the component value of the vector of user u_i can be calculated as :

$$p(d_\alpha|u_i) = \frac{p(u_i|d_\alpha)p(d_\alpha)}{p(u_i)} \sim p(u_i|d_\alpha)p(d_\alpha) \tag{8}$$

Since $\sum_{\alpha=1}^{D} p(d_\alpha|u_i) = 1$, we are able to calculate $p(d_\alpha|u_i)$ by the probabilities obtained in EM methods. $p(d_\alpha|u_i)$ measures how the interests of u_i relate to the category of knowledge in the dimension α.

In each iteration, the time complexity of the above EM algorithm is $O(C * D)$, which is linear to both the size of the annotations and the size of the concept space dimension. Notice that the co-occurrence number is usually much larger than any one data set of entities, so the indicator matrix I occupies most of the storage spaces. We interleave the output of E-step and the input of M-step without saving indicator matrix I. Hence the space complexity without the storage of raw triples in the algorithm is $O(D*(K+M+N))$.

3.2 Experiments

We collected a sample of Del.icio.us data by crawling its website during March 2005. The data set consists of 2,879,614 taggings made by 10,109 different users on 690,482 different URLs with 126,304 different tags. In our experiments, we reduced the raw data by filtering out the users who annotate less than 20 times, the URLs annotated less than 20 times, and the tags used less than 20 times. The experiment data contains 8676 users, 9770 tags and 16011 URLs. Although it is much less than the raw data, it still contains 907,491 triples. We perform EM iterations on this data set. Figure 2 presents the log-likelihood on the social annotations data by choosing different number of dimensions and with different iteration times.

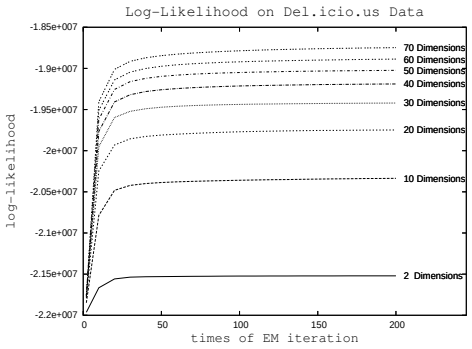

Figure 2: The Log-Likelihood on the times of iteration of different number of aspects

Table 1: Top 5 tags in 10 out of 40 conceptual dimensions

1	java programming Java eclipse software
2	css CSS web design webdesign
3	blog blogs design weblogs weblog
4	music mp3 audio Music copyright
5	search google web Google tools
6	python programming Python web software
7	rss RSS blog syndication blogs
8	games fun flash game Games
9	gtd productivity GTD lifehacks organization
10	programming perl development books Programming

In Figure 2, we can find that the log-likelihood increases very fast from 2-dimensions to 40-dimensions and slows down in dimensions higher than 40. Because the web bookmarks collected on Del.icio.us are mainly in the field of IT, the knowledge repository is relatively small and the conceptual space with 40 dimensions is basically enough to represent the major category of meanings in Del.icio.us. Higher dimensions are very probably redundant dimensions which can be replaced by others or a combination of other dimensions. Large number of dimensions may also bring out the problem of over-fitting. As to iteration, iterate 80 times can provide satisfying result and more iterations won't give great improvement and may cause over-fitting. In our experiment, we model our data with 40 dimensions and calculate the parameters by iterating 80 times.

We choose the top 5 tags according to $p(t_k|d_\alpha)$ on each dimension, and randomly list 10 dimensions in Table 1. From this table, we can find that each dimension concern with a special category of semantics. Dimension 1 is mainly about 'programming', and dimension 5 talk about 'search engines'. The semantically related tags have high component values in the same dimension, such as 'mp3' and 'music', while 'css' and 'CSS', 'games' and 'Games' are actually about the same thing.

We also study the ambiguity of different tags on dimensions. The entropy of a tag can be computed as

Table 2: Tags and their entropy

NO.	Tags	Entropy	Tags	Entropy
1	todo	3.08	cooking	0
2	list	2.99	blogsjava	0
3	guide	2.92	nu	0
4	howto	2.84	eShopping	0
5	online	2.84	snortgiggle	0
6	tutorial	2.78	czaby	0
7	articles	2.77	ukquake	0
8	collection	2.76	mention	0
9	the	2.71	convention	0
10	later	2.70	wsj	0

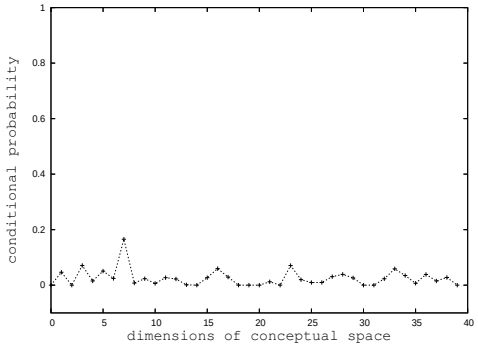

Figure 3: Conditional Distribution of Tag 'todo' on dimensions of conceptual space

$$E = -\sum_{\alpha=1}^{D} p(d_\alpha|t) \log p(d_\alpha|t) \tag{9}$$

and it can be used as an indicator of the ambiguities of the tag. The top 10 and bottom 10 tags of ambiguity in our experiment are shown in Table 2. We find that the tag 'todo' in Figure 3 has the highest entropy. It's the most ambiguous tag used in Del.icio.us and its distribution on dimensions are very even. The tag 'cooking' in Figure 4 has the lowest entropy. Its meaning is quite definite in this social annotation data set. We will take a looking at the tag 'xp' in Figure 5, which has 2 comparatively high components in dimension 27

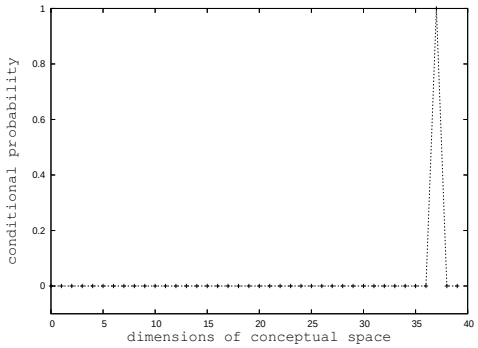

Figure 4: Conditional Distribution of Tag 'cooking' on dimensions of conceptual space

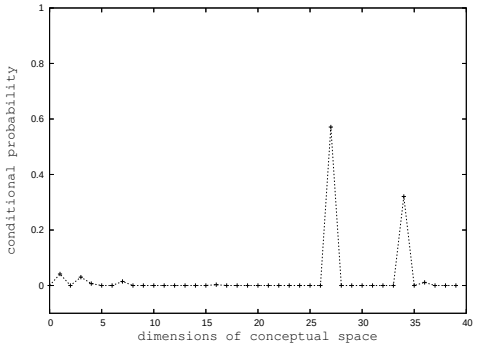

Figure 5: Conditional Distribution of Tag 'xp' on dimensions of conceptual space

and 34 while keeps very low on other dimensions. The top 5 tags on dimension 27 are "security windows software unix tools", on dimension 34 they are "java programming Java eclipse software" . The word 'xp' can be an abbreviation of two phrases. One is 'Window XP' which is an operating system. The other is 'Extreme Programming' which is a software engineering method. Many extreme programming toolkits are developed by 'Java' in 'Eclipse' IDE. In this case, the vector representation of the tag 'XP' identifies its meaning very clearly through its coordinates in the conceptual space. Similar results can be achieved for resources and users. This enables us to to give semantic annotation to users, tags and resources in the form of vectors, which can represent their meanings in the conceptual space. For tags, annotations identify the ambiguity and synonymy; For users, annotation will present the users' interests which can be utilized for personalized search; For web resources, annotation can present the semantics of contents in the resources.

3.3 Semantic Search and Discovery

After deriving the emergent semantics from social annotations, the semantics of user interests, tags and web resources can be represented by vectors in the conceptual space. Based on these semantic annotations, an intelligent semantic search system can be implemented. In such a system, users can query with a boolean combination of tags and other keywords, and obtain resources ranked by relevance to users' interests. If the meaning of input query is ambiguous, hints will be provided for a more detailed search on a specific meaning of a tag.

3.3.1 Basic Search Model

In this part, we develop the basic search model. Advanced functions such as personalized search and complicated query support are built upon it. The basic model deals with queries that are a single tag and rank semantic related resources without considering personalized information of the user. This problem can be converted to a probability problem.

$$p(r|t) = \sum_{\alpha=1}^{D} p(r|d_\alpha) p(d_\alpha|t) \tag{10}$$

In (10), the effects of all dimensions are combined together to generate the conditional probability. The return resources will be ranked by the conditional probability $p(r|t)$.

We can also provide a more interactive searching inter-

face, when a user queries with tag t_j which is ambiguous and have a high entropy calculated in (9) larger than a predefined threshold. The user will, in addition to the usual query results, also get a list of categories of knowledge with top tags as further disambiguation choices for the tag. The categories are ranked by $p(d|t_j)$. When the user chooses a specific category of knowledge, the resources will return ranked by $p(u|d)$, which helps to narrow the search scope and increase search accuracy.

3.3.2 Resource Discovery

The basic search model developed above searches and ranks related resources of a given tag according to the conditional probability $p(r|t)$, which is directly related to the similarity of their vectors in the conceptual space. This model is thus totally based on the emergent semantics of social annotations without using any keyword matching metrics. We can go into this direction even further by discovering highly semantically related resources which are even not tagged by the query tag by any user before. We can extend our basic model to support this if we force:

$$p(r|t) = \begin{array}{cc} \sum_{\alpha=1}^{D} p(r|d_\alpha)p(d_\alpha|t) & : \quad n_{tr} = 0 \\ 0 & : \quad n_{tr} > 0 \end{array} \tag{11}$$

In (11), n_{tr} denotes the number of co-occurrences of the tag and resource. We filter out the already-tagged resources by set their conditional probability to zero and only return resources that are not tagged by the query tag and rank them by $p(r|t)$. We implemented this resource discovery search on the Del.icio.us data set and it produces interesting results. For example, when a user searches with the tag 'google' in this resource discovery mode, the returned URL list contains an introduction of 'Beagle' which is a desktop search tool for GNOME on linux. This web page is never tagged by 'google' by any user in the data set. It even does not contain the word 'google' in its web page content. This page thus can not be found using traditional search methods, such as keyword search or search based on tags, although 'beagle' and 'google' are semantically related. More interestingly, if queryed with 'delicious', the method will return web pages that are highly related to semantic web technologies such as RDF and FOAF. This search result reveals interesting semantic connection between the Del.icio.us web site and the semantic web. We list these two discovery results of 'delicious' and 'google' in appendix section A.

3.3.3 Personalized Search Model

Due to the diversity of users in the social bookmarking service, it's possible for two users to search with the same tag but demand different kinds of resources. For example, searching with the tag "xp", a programmer may prefer resources related with "Extreme Programming" while a system manager may want to know about the operating system "Window XP". Since users' interests can be represented by vectors in the conceptual space, we can attack the problem by integrating personalized information in the semantic search. It can be formalized by:

$$\begin{aligned} p(r|u,t) &= \sum_{\alpha}^{D} p(r|d_\alpha)p(d_\alpha|u,t) \\ &= \sum_{\alpha}^{D} p(r|d_\alpha)\frac{p(u,t|d_\alpha)p(d_\alpha)}{p(u,t)} \\ &\sim \sum_{\alpha}^{D} p(r|d_\alpha)p(u|d_\alpha)p(t|d_\alpha)p(d_\alpha) \end{aligned} \tag{12}$$

In our model, as shown in Figure 1, entities can be viewed independently in the conceptual space, thus $p(u,t|d_\alpha) = p(u|d_\alpha)p(t|d_\alpha)$. $p(u,t)$ keeps the same in one search process, and $\sum_{j=1}^{N} p(r_j|u,t) = 1$, so we can calculate the resources' semantic relatedness $p(r|u,t)$ by (12).

3.3.4 Complicated Query Support

In the above model, users can only query with a single tag. That's far from enough to express complicated query requirements. If the web resources are documents, users may want to search its contents using keywords in addition to tags. We extend our basic model to support queries that can be a boolean combination of tags and other words appearing in the resources. Let q denote the complicated query. The basic model can be modified to (13).

$$p(r|q) \sim \sum_{\alpha=1}^{D} p(r|d_\alpha)p(q|d_\alpha)p(d_\alpha) \tag{13}$$

Now the problem turns to estimate $p(q|d_\alpha)$. Let's start from the simplest case. Suppose the query q is a single word w in a document and is not a tag. We utilize the document resources as an intermediate, and convert the problem to estimate $p(w|r)$ in (14).

$$p(w|d_\alpha) = \sum_{j=1}^{N} p(w|r_j)p(r_j|d_\alpha) \tag{14}$$

$p(w|r_j)$ can be viewed as the probability of producing a query word w from the corresponding language model of the document resource r_j. We can use the popular Jelinek-Mercer [16] language model to estimate $p(w|r_j)$.

$$p(w|r_j) = (1-\lambda)p_{ml}(w|r_j) + \lambda p(w|COL) \tag{15}$$

where $p_{ml}(w|r_j) = \frac{c(w,r_j)}{\sum_w c(w,r_j)}$. $c(w,r_j)$ denotes the count of word w in resource document r_j. $p(w|C)$ is the general frequency of w in the resource document collection COL.

When the input query q is a boolean combination of tags and other words, we adopt the extended retrieval model [28] to estimate $p(q|d)$. The query is represented in the following manner:

$$q = \{k_1 : a_1, k_2 : a_2, \ldots, k_p : a_p\} \tag{16}$$

In (16), k_i denote the ith component in the query, which can be either a tag or a keyword. a_i denote the weight of the component k_i in the query, which measures the importance of this component in the query. In our experiments, we assigned equal weights to each component. p is the number of components. The boolean combination of these components could be either 'and' or 'or'. The probability of 'and' query and 'or' query can be calculated in (17) and (18) respectively using [28].

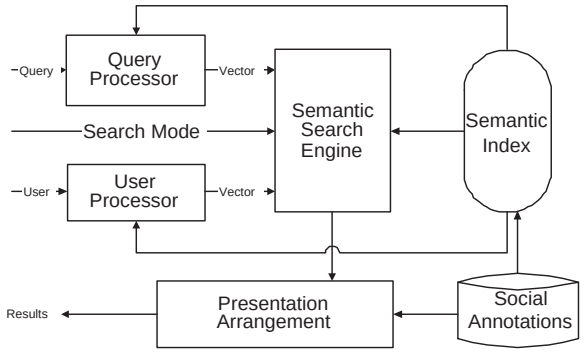

Figure 6: The framework of our social semantic search system

$$p(q_{and}|d) = 1-[\frac{a_1^p(1-p(k_1|d))^p + \ldots + a_n^p(1-p(k_n|d))^p}{a_1^p + a_2^p + \ldots + a_n^p}]^{\frac{1}{p}} \quad (17)$$

$$p(q_{or}|d) = [\frac{a_1^p p(k_1|d)^p + a_2^p p(k_2|d)^p + \ldots + a_n^p p(k_n|d)^p}{a_1^p + a_2^p + \ldots + a_n^p}]^{\frac{1}{p}} \quad (18)$$

For more complicated boolean combinations that contains both 'and' and 'or', it can be calculated using (17) and (18) recursively. For example, the query $\{(t_A : 0.3\ and\ w_A : 0.4) : 0.2\ or\ (t_B : 0.1)\}$ in which t_A and t_B are tags while w_A is a keyword but is not tag. We first calculate the 'and' probability of t_A and w_A,

$$p(t_A\ and\ w_A|d) = 1-\sqrt{\frac{0.3^2(1-p(t_A|d))^2 + 0.4^2(1-p(w_A|d))^2}{0.3^2+0.4^2}}$$

and then calculate the total conditional probability.

$$p(q|d) = \sqrt{\frac{0.2^2 p(t_A\ and\ w_A|d)^2 + 0.1^2 p(t_B|d)^2}{0.2^2+0.1^2}}$$

$p(t_A|d)$ and $p(t_B|d)$ are acquired after the EM iterations and $p(w_A|d)$ is calculated in (14).

Our search models are quite flexible. The web bookmarks discovery model, personalized search model and complicated query support model are independent optional parts built on the basic model. We can use them separately or combine several of them together. For example, (19) combined all of them together.

$$p(r|u,q) \sim \sum_{\alpha=1}^{D} p(r|d_\alpha)p(u|d_\alpha)p(q|d_\alpha)p(d_\alpha) \quad (19)$$

4. IMPLEMENTATION AND EVALUATION

In this section, we describe the implementation of a semantic search and discovery system[2] based on the models developed above, and the application of this system to the Del.icio.us social annotations data. Figure 6 shows the framework of our system, which can be divided into two parts by function. The back-end part collects and builds semantic index on folksonmies data while the foreground accepts query, retrieve related resources and present results in a friendly manner.

In the back-end part, after the data is collected and stored to the 'Social Annotations DB', the system will start to run the EM algorithm with respect to the tripartite model developed in Section 3.1 and compute the vectors of users, web resources and tags as the semantic index. For the words which are not tags but appear in the web pages of URLs, a language model approach developed in Section 3.3.4 is implemented to index them.

In the foreground part, when a user initiates a search action, three parameters are passed to the system: the input query, user's identification and the search model (personalized or discovery or both). In the 'query processor' unit, the input query q is first parsed to a boolean combination of tags and other keywords and then mapped to a vector

$$\langle p(q|d_1), p(q|d_2), \ldots, p(q|d_D)\rangle$$

according to the method introduced in Section 3.3.4. In the 'user processor' unit, the user will be identified and mapped to the related vector stored in the 'semantic index' unit. The search engine receives the output vectors of query processor and user process, finds the related URLs according to the input search mode, and then passes the raw result to the 'presentation arrangement' unit, where the results are refined to provide an interactive web user-interface.

One important difference of our search model is the ability to discover semantically-related web resources from emergent semantics, even if the web resource is not tagged by the query tags and does not contain query keywords. This search capability is not available in the current social bookmarking services. We evaluate the effectiveness of this discovery ability using our implementation system.

We choose 5 widely used tags 'google', 'delicious', 'java', 'p2p' and 'mp3' on Del.icio.us folksonomy data set, and separately input them into our system. The system works in the resources discovery mode (filtering out the URLs tagged by these tags), and returns the discovered list of URLs. We choose top 20 URLs in every list to evaluate the semantic relatedness between the tags and the results. As the URLs in Del.icio.us are mainly on the IT subjects, we invited 10 students in our lab who are doctor or master candidates majoring in computer science and engineering to take part in the experiment. Each student is given all the 100 URLs. They are asked to judge the semantic relatedness between the tag and the web pages of URLs based on their knowledge and score the relatedness from 0 point (not relevant) to 10 points (highly relevant). We average their scores on each URL and use the graded precision to evaluate the effectiveness of the resources discovery capability. The graded precision is:

$$gp_i = \frac{\sum_{\alpha=1}^{i} score(\alpha)}{i*10} : i <= 20 \quad (20)$$

In (20), $score(\alpha)$ denotes the average score of the αth URL for a tag search. For each tag search, we calculate gp_i, with i ranging from 1 to 20 to represent the top i results. The graded precision result is shown in Figure 7.

5. RELATED WORK

Since it's a quite new service and topic, there are only very few published studies on social annotations. [10] gives a detailed analysis of the social annotations data in Del.icio.us

[2]The system can be accessed via http://apex.sjtu.edu.cn:50188

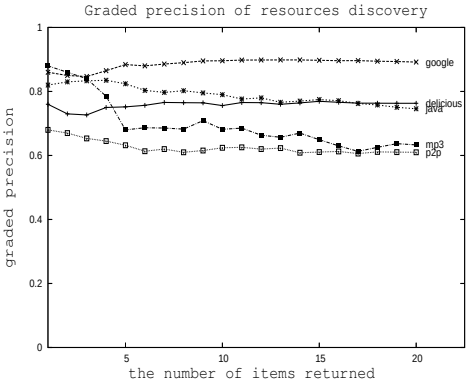

Figure 7: The graded precision

from both the static and dynamic aspects. They didn't, however, make deep analysis on the semantics of these annotations. [25] proposes to extend the traditional bipartite model of ontology with a social dimension. The author found the semantic relationships among tags based on their co-occurrences with users or resources but without considering the ambiguity and group synonymy problems. It also lacks a method to derive and represent the emergent semantics for semantic search.

Semantic annotation is a key problem in the Semantic Web area. A lot of work has been done about the topic. Early work like [26, 31] mainly uses an ontology engineering tool to build an ontology first and then manually annotate web resources in the tool. In order to help automate the manual process, many techniques have been proposed and evaluated. [7] learns from a small amount of training examples and then automatically tags concept instances on the web. The work has been tested on a very large-scale basis and achieves impressive precision. [4] helps users annotate documents by automatically generate natural language sentences according to the ontology and let users interact with these sentences to incrementally formalize them. Another interesting approach is proposed by [5] that utilizes the web itself as a disambiguation source. Most annotations can be disambiguated purely by the number of hits returned by web search engines on the web. [6] improves the method using more sophisticated statistical analysis. Given that many web pages nowadays are generated from a backend database, [12] proposes to automatically produce semantic annotations from the database for the web pages. Information extraction techniques are employed by [8] to automatically extract instances of concepts of a given ontology from web pages. However, this work on semantic annotations follows the traditional top-down approach to semantic annotation which assumes that an ontology is built before the annotation process.

Much work has been done to help users manage their bookmarks on the (semantic) web such as [17]. [11] gives a good review of the social bookmarks tools available. These tools help make the social bookmarking easy to use but lack capabilities to derive emergent semantics from the social bookmarks.

Work on emergent semantics [19, 2] has appeared recently, for example [35, 1, 15]. [1] proposes an emergent semantics framework and shows how the spreading of simple ontology mappings among adjacent peers can be utilized to incrementally achieve a global consensus of the ontology mapping. [15] described how to incrementally obtain a unified data schema from the users of a large collection of heterogeneous data sources. [35] is more related to our work. It proposes that the semantics of a web page should not and cannot be decided alone by the author. The semantics of a web page is also determined by how the users use the web page. This idea is similar to our thought. In our work, a URL's semantics is determined from its co-occurrences with users and tags. However, our method of achieving emergent semantics is different from [35]. We use a probabilistic generative model to analyze the annotation data while [35] utilizes the common sub-paths of users' web navigation paths.

6. CONCLUSIONS AND FUTURE WORK

Traditional top-down approach to semantic annotation in the Semantic Web area has a high barrier to entry and is difficult to scale up. In this paper, we propose a bottom-up approach to semantic annotation of the web resources by exploiting the now popular social bookmarking efforts on the web. The informal social tags and categories in these social bookmarks is coined a name called "folksonomy". We show how a global semantic model can be statistically inferred from the folksonomy to semantically annotate the web resources. The global semantic model also helps disambiguate tags and group synonymous tags together in concepts. Finally, we show how the emergent semantics can be used to search and discover semantically-related web resources even if the resource is not tagged by the query tags and does not contain any query keywords.

Unlike traditional formal semantic annotation based on RDF or OWL, social annotation works in a bottom-up way. We will study the evolution of social annotations and its combination with formal annotations. For example, enrich formal annotations with social annotations.

Social annotations are also sensitive to the topic drift in the user community. With the increasing of a special kind of annotations, the answers for the same query may change. Our model can reflect this change but requires re-calculation on the total data set periodically which is quite time consuming. One goal of our future work is to improve our model to support incremental analysis of the social annotations data.

7. ACKNOWLEDGEMENT

The authors would like to thank IBM China Research Lab for its continuous support and cooperation with Shanghai JiaoTong University on the Semantic Web research.

8. REFERENCES

[1] K. Aberer, P. Cudre-Mauroux, and M. Hauswirth. The chatty web: Emergent semantics through gossiping. In *Proc. of 12th Intl. Conf. on World Wide Web (WWW2003)*, 2003.

[2] K. Aberer and et al. Emergent semantics principles and issues. In *Proc. of Database Systems for Advanced Applications*, LNCS 2973, 2004.

[3] T. Berners-Lee, J. Hendler, and O. Lassila. The Semantic Web. *Scientific American*, 284(5):34–43, May 2001.

[4] J. Blythe and Y. Gil. Incremental formalization of document annotations through ontology-based paraphrasing. In *Proc. of the 13th Conference on*

World Wide Web (WWW2004), pages 455–461. ACM Press, 2004.

[5] P. Cimiano, S. Handschuh, and S. Staab. Towards the self-annotating web. In *Proc. of the 13th Intl. World Wide Web Conference (WWW2004)*, 2004.

[6] P. Cimiano, G. Ladwig, and S. Staab. Gimme the context: Context-driven automatic semantic annotation with C-PANKOW. In *Proc. of the 14th Intl. World Wide Web Conference (WWW2005)*, 2005.

[7] S. Dill, N. Eiron, D. Gibson, D. Gruhl, R.Guha, A. Jhingran, T. Kanungo, S. Rajagopalan, A. Tomkins, J. A.Tomlin, and J. Y.Zien. SemTag and Seeker: Bootstrapping the semantic web via automated semantic annotation. In *Proc. of the 12th Intl. World Wide Web Conference (WWW2003)*, pages 178–186, 2003.

[8] O. Etzioni, M. Cafarella, D. Downey, S. Kok, A.-M. Popescu, T. Shaked, S. Soderland, D. S.Weld, and A. Yates. Web-scale information extraction in KnowItAll (preliminary results). In *Proc. of the 13th Intl. World Wide Web Conf.(WWW2004)*, 2004.

[9] N. F. C. N. Pereira and L. Lee. Distributional clustering of English words. In *Proceedings of the Association for Computational Linguistics*, pages 183–190, 1993.

[10] S. A. Golder and B. A. Huberman. The structure of collaborative tagging systems. http://www.hpl.hp.com/research/idl/papers/tags/, 2005.

[11] T. Hammond, T. Hannay, B. Lund, and J. Scott. Social bookmarking tools (i) - a general review. *D-Lib Magazine*, 11(4), 2005.

[12] S. Handschuh, S. Staab, and R. Volz. On deep annotation. In *Proc. of the 12th Intl. World Wide Web Conference (WWW2003)*, pages 431–438, 2003.

[13] T. Hofmann. Probabilistic latent semantic indexing. In *Proc. of the 22nd ACM SIGIR Conference*, 1999.

[14] T. Hofmann and J. Puzicha. Statistical models for co-occurrence data. Technical report, A.I.Memo 1635, MIT, 1998.

[15] B. Howe, K. Tanna, P. Turner, and D. Maier. Emergent semantics: Towards self-organizing scientific metadata. In *Proc. of the 1st Intl. IFIP Conference on Semantics of a Networked World: Semantics for Grid Databases (ICSNW 2004)*, LNCS 3226, 2004.

[16] F. Jelinek and R. L. Mercer. Interpolated estimation of Markov source parameters from sparse data. In *Proceedings of Workshop on Pattern Recognition in Practice*, 1980.

[17] J. Kahan, M.-R. Koivunen, E. Prud'Hommeaux, and R. R. Swick. Annotea: An open RDF infrastructure for shared web annotations. In *Proc. of the 10th Intl. World Wide Web Conference*, 2001.

[18] A. Kiryakov, B. Popov, D. Ognyanoff, D. Manov, A. Kirilov, and M. Goranov. Semantic annotation, indexing, and retrieval. In *Proc. of the 2nd Intl. Semantic Web Conference (ISWC2003)*, 2003.

[19] A. Maedche. Emergent semantics for ontologies. *IEEE Intelligent Systems*, 17(1), 2002.

[20] F. Manola and E. Miller. RDF Primer. *W3C Recommendation*, 2004.

[21] A. Mathes. Folksonomies - cooperative classification and communication through shared metadata. Computer Mediated Communication, LIS590CMC (Doctoral Seminar), Graduate School of Library and Information Science, University of Illinois Urbana-Champaign, December 2004.

[22] R. M. Bada, D. Turi and R. Stevens. Using reasoning to guide annotation with gene ontology terms in goat. *SIGMOD Record (Special issue on data engineering for the life sciences)*, June 2004.

[23] D. L. McGuinness and F. van Harmelen. OWL Web ontology language overview. *W3C Recommendation*, 2004.

[24] P. Merholz. Metadata for the masses. http://www.adaptivepath.com/publications/essays/archives/000361.php, accessed at May, 2005, October 2004.

[25] P. Mika. Ontologies are us: A unified model of social networks and semantics. In *Proc. of 4rd Intl. Semantic Web Conference (ISWC2005)*, 2005.

[26] N. F. Noy, M. Sintek, S. Decker, M. Crubezy, R. W. Fergerson, and M. A. Musen. Creating semantic web contents with Protege-2000. *IEEE Intelligent Systems*, 2(16):60–71, 2001.

[27] K. Rose. Deterministic annealing for clustering, compression. *Proceedings of the IEEE*, 86(11), 1998.

[28] G. Salton, E. A. Fox, and H. Wu. Extended boolean information retrieval. *Communications of the ACM*, 26(11):1022–1036, 1983.

[29] J. Schachter. Del.icio.us about page. http://del.icio.us/doc/about, 2004.

[30] G. L. S. Deerwester, S. T. Dumais and R. Harshman. Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, 1990.

[31] S. Handschuh and S. Staab. Authoring and annotation of web pages in CREAM. In *Proc. of the 11th Intl. World Wide Web Conference (WWW2002)*, 2002.

[32] G. Smith. Atomiq: Folksonomy: social classification. http://atomiq.org/archives/2004/08/folksonomy_social_classification.html, Aug 2004.

[33] D. Song and P. Bruza. Discovering information flow using a high dimensional conceptual space. In *Proceedings of the 24th International ACM SIGIR Conference*, pages 327–333, 2001.

[34] J. Udell. Collaborative knowledge gardening. *InfoWorld, August 20*, August 2004.

[35] W. I. Grosky, D. V. Sreenath, and F. Fotouhi. Emergent semantics and the multimedia semantic web. *SIGMOD Record*, 31(4), 2002.

[36] L. Zhang, X. Wu, and Y. Yu. Emergent semantics from folksonomies, a quantitative study. *Special issue of Journal of Data Semantics on Emergent Semantics, to appear*, 2006.

APPENDIX

A. RESOURCES DISCOVERY

A.1 Discovery results for query tag 'delicious'

1	http://www.betaversion.org/ stefano/linotype/news/57
2	http://www.amk.ca/talks/2003-03/
3	http://www.ldodds.com/foaf/foaf-a-matic.html
4	http://www.foaf-project.org/
5	http://gmpg.org/xfn/
6	http://www.ilrt.bris.ac.uk/discovery/rdf/resources/
7	http://xml.mfd-consult.dk/foaf/explorer/
8	http://xmlns.com/foaf/0.1/
9	http://simile.mit.edu/welkin/
10	http://www.xml.com/pub/a/2004/09/01/ hack-congress.html
11	http://www.w3.org/2001/sw/
12	http://simile.mit.edu/
13	http://jena.sourceforge.net/
14	http://www.w3.org/RDF/
15	http://www.foafspace.com/

A.2 Discovery results for query tag 'google'

1	http://www.musicplasma.com/
2	http://www.squarefree.com/bookmarklets/
3	http://www.kokogiak.com/amazon4/default.asp
4	http://www.feedster.com/
5	http://http://www.gnome.org/projects/beagle/
6	http://www.faganfinder.com/urlinfo/
7	http://www.newzbin.com/
8	http://www.daypop.com/
9	http://www.copernic.com/
10	http://www.alltheweb.com/
11	http://a9.com/-/search/home.jsp?nocookie=1
12	http://snap.com/index.php/
13	http://www.blinkx.tv/
14	http://www.kartoo.com/
15	http://www.bookmarklets.com/