Community-Based Ranking of the Social Web

Said Kashoob James Caverlee Krishna Kamath
Department of Computer Science and Engineering
Texas A&M University
College Station, TX 77843 USA
{kashoob,caverlee,krishna}@cse.tamu.edu

ABSTRACT

The rise of social interactions on the Web requires developing new methods of information organization and discovery. To that end, we propose a generative community-based probabilistic tagging model that can automatically uncover communities of users and their associated tags. We experimentally validate the quality of the discovered communities over the social bookmarking system Delicious. In comparison to an alternative generative model (Latent Dirichlet Allocation (LDA), we find that the proposed community-based model improves the empirical likelihood of held-out test data and discovers more coherent interest-based communities. Based on the community-based probabilistic tagging model, we develop a novel community-based ranking model for effective community-based exploration of socially-tagged Web resources. We compare community-based ranking to three state-of-the-art retrieval models: (i) BM25; (ii) Cluster-based retrieval using K-means clustering; and (iii) LDA-based retrieval. We find that the proposed ranking model results in a significant improvement over these alternatives (from 7% to 22%) in the quality of retrieved pages.

Categories and Subject Descriptors

H.3.3 [**Information Search and Retrieval**]: Clustering and Retrieval models

General Terms

Algorithms

Keywords

Community, Ranking, Social, Tagging

1. INTRODUCTION

The past few years have seen the rapid rise of all things "social" on the web – from the growth of online social networks like Facebook, to user-contributed content sites like Flickr and YouTube, to social bookmarking services like Delicious. Whereas traditional approaches to organizing and accessing the Web's massive amount of information have focused on *content-based* and *hyperlink-based* approaches (e.g., PageRank [33], HITS [22]), these social systems offer rich opportunities for *user-based* exploration and analysis of the Web by building on the unprecedented access to the interests and perspectives of millions of users.

In this paper, we examine a popular and growing type of *user-based* social system, namely *social bookmarking systems*. These bookmarking systems allow users to personally organize web pages, images, videos, and other web media by tagging (or annotating) each resource with simple keywords or phrases. For example, a user could tag the CNN homepage (`http://www.cnn.com`) with the tags `news`, `politics`, and `cnn`. Individually, each user is now explicitly linked to a Web resource by a tag or group of tags (see Figure 1). Collectively, the social bookmarking system can aggregate thousands of user's perspectives on each tagged resource to enrich the resource with large-scale socially-generated metadata. The scale of these systems is large; indeed, a recent study by Heymann et al. found that the single bookmarking site Delicious has already led to the tagging of over 150 million web pages [19]. Based on the scale of this fundamentally *user-based* effort, there has been growing excitement at augmenting traditional content-based and hyperlink-based web search and browsing through the incorporation of tag information from social bookmarking services, e.g., [3], [6], [19], [24], [41], [42], and [43].

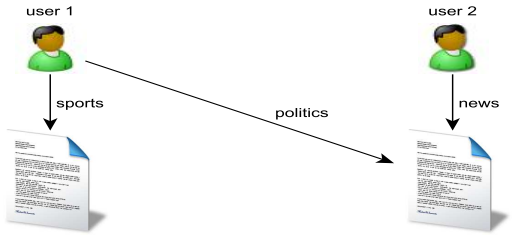


Figure 1: Two users linking to Web resources

Our particular interest in this paper is on the impact of *community* in these social bookmarking systems. The notion of community is fundamental to the Social Web – be it friendships on Facebook, groups of similarly-interested users who comment on YouTube videos, collections of Wikipedia contributors who specialize in certain topics, and so on. Social bookmarking systems, however, aggregate what would appear to be the independent and uncoordinated tagging actions of a large and heterogeneous tagger population, meaning that it is not obvious that communities of users exist or are detectable. Given the strong evidence of community in the other areas of the Social Web and recent research

HT'10, June 13–16, 2010, Toronto, Ontario, Canada..

that has indicated the evidence of coherent tag-based and resource-based clusters in social bookmarking systems [35, 8, 41, 44, 34], we are interested to explore: (i) if *user-based communities* manifest themselves in social bookmarking systems and how to model them; and (ii) whether this community-based perspective can enhance how users explore the web of socially tagged resources.

Concretely, we first posit that the observed tagging information in a social bookmarking system (e.g., that User A applied the tag `politics` to the resource `http://www.cnn.com`, that User B applied the tag `sports` to the resource `http://www.espn.com`, and so on) is the product of an underlying community structure, in which users belong to *implicit groups of interest* (e.g., students, sports fans). Our hope is that by analyzing the tag-based linking structure, we can uncover these implicit communities without any a priori knowledge of users and their affiliated communities (see Figure 2).

If these implicit communities can be extracted from large-scale social bookmarking services, it may be advantageous to leverage this community structure to enhance user-based exploration over the web of socially tagged resources. Instead of guiding users to resources that a user already knows about (via his own tags) or that are globally well-known (e.g., a web page that many other users have already tagged), we seek to develop new community-based exploration approaches that emphasize the community's implicit view (e.g., to identify resources that are relevant to the implicit *sports* community). Concretely, we propose a novel community-based ranking model that is designed with this intuition in mind.

To summarize, the contributions of this paper are:

- We propose a generative community-based probabilistic tagging model in which users belong to implicit groups of interest (e.g., students, sports fans) and probabilistically select tags with which to bookmark resources. Coupled with Gibbs sampling parameter estimation, the community-based model can automatically uncover these communities of implicitly related users and their associated tags.
- We experimentally validate the quality of the discovered communities over the social bookmarking system Delicious. In comparison to an alternative generative model (Latent Dirichlet Allocation (LDA) [5]), we find that the proposed community-based model improves the empirical likelihood of held-out test data and discovers more coherent interest-based communities.
- Based on the community-based probabilistic tagging model, we develop a novel community-based ranking model for effective community-based exploration of socially-tagged Web resources. The ranking model leverages the discovered community structure to implicitly connect user, tags, and resources for more effective information exploration and discovery.
- We compare community-based ranking to three state-of-the-art retrieval models: (i) BM25; (ii) Cluster-based retrieval using K-means clustering; and (iii) LDA-based retrieval. We find that the proposed ranking model results in a significant improvement over these alternatives (from 7% to 22%) in the quality of retrieved pages.

The rest of the paper is organized as follows. Section 2 introduces the overall intuition and preliminary definitions for community-based modeling and ranking. In Section 3, we describe the probabilistic community model and how it can be used to uncover implicit communities of users and their tags. Based on this model, we propose a novel community-based ranking approach in Section 4. Section 5 presents experimental evidence over Delicious, and we conclude in Section 6 with our final thoughts.

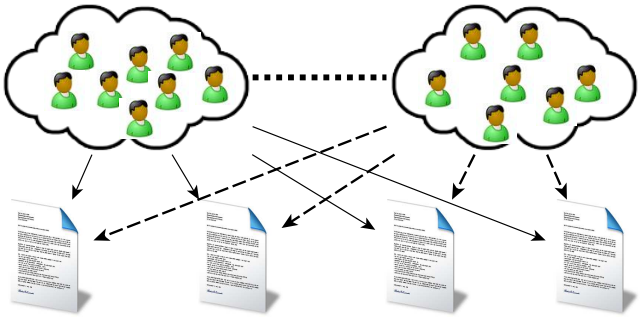


Figure 2: Community-based tagging

2. OVERVIEW

Formally, we consider a universe of discourse $\mathcal{U}$ consisting of R resources (e.g., images, videos, Web pages, ...), U users, and a vocabulary of T tags. A user u_i is linked to a resource r_j by the tag t_k used to bookmark the resource. This linkage is represented as the tuple (u_i, r_j, t_k). The notation introduced here and throughout the paper is summarized in Table 1. Note that users can tag a single resource with multiple tags. Figure 1 illustrates a simple universe composed of two users, two resources, and three tags. In practice, popular social bookmarking services include Delicious for annotating web pages, Flickr for annotating images, and CiteULike for annotating scholarly articles, among many others.

Although these bookmarking services attract a wide range of taggers – who may vary greatly in interest, expertise, and language – we hypothesize that there are implicit groups of users formed around a common community perspective. For example, an image of a Tyrannosaurus rex could be bookmarked by users belonging to the underlying scientist community (e.g., with tags like `cretaceous` and `theropod`), and by users belonging to the underlying elementary student community (e.g., with tags like `t-rex` and `meat-eater`). These communities are not explicitly declared nor even obvious to members of each community; for example, two elementary students may belong to the same community through their use of common tags or through their tagging of common resources, even without knowledge of each other.

This hypothesis of implicit community structure is motivated by previous studies of social bookmarking systems. For example, Golder and Huberman [15] found a number of clear structural patterns in Delicious, including the stabilization of tags over time, even in the presence of a large and heterogeneous user population. This stabilization (which might be counter-intuitive, especially in contrast to the tightly controlled metadata produced by domain experts) suggests a shared knowledge in bookmarking systems. These results are echoed by Halpin et al. [16], who found a power-law distribution for Delicious tags applied to web pages – meaning that in the aggregate, distinct users independently described a page using a common tagging vocabulary. Similar results can be found elsewhere, including [9], [7], and [37]. Along this line, there have been some recent efforts to identify clusters (topics) of tags and resources, including: [26], which mined tag-based topics via association rules; [42], which iteratively determines user interests and resource topics over a bipartite graph where users and resource are nodes and tagging counts are edges; and [35], which identifies groups of web resources by clustering them via their tags using both K-means and probabilistic clustering. These results motivate our interest in uncovering hidden communities that could help explain these phenomena.

Table 1: Notation

symbol	Description	symbol	Description	symbol	Description
$\mathcal{U}$	corpus	S	social tagging document	θ	A resource's community distribution
U	user vocabulary	z	topic	ϕ	A topic's tag distribution
T	tag vocabulary	c	community	τ	A community's user distribution
N_i	document length	L	number of communities	α	user hyperprior
r	resource	R	number of documents	β	community hyperprior
t	tag	u	user	γ	tag hyperprior

Toward the goal of identifying community in social bookmarking systems, we posit the existence of L communities that are implicit in the universe of discourse $\mathcal{U}$, where each community is composed of users and tags that are representative of the community's perspective. Since community membership is not fixed, we model membership as a probability distribution, where each user has some probability of belonging to any community.

[Definition] Social Tagging Community: *A social tagging community c is composed of (i) a probability distribution over users in U such that $\sum_{u \in U} p(u|c) = 1$, where $p(u|c)$ indicates membership strength for each user u in community c; and (ii) a probability distribution over tags in the vocabulary T such that $\sum_{t \in T} p(t|c) = 1$, where $p(t|c)$ indicates membership strength for each tag t in community c.*

In practice, communities are *hidden* from us; all we may observe are the user, resource, tag tuples (u_i, r_j, t_k) that are the result of these communities and the categories they have selected. The challenge that we address in the next section is how to recover these hidden communities from the observable tagging data. Based on the proposed community discovery algorithm, we continue in the following section with an investigation of how to support community-driven ranking over web resources.

3. MODELING AND UNCOVERING SOCIAL TAGGING COMMUNITIES

In this section, we present our study of community modeling over social tagging data. Our goal is to identify social tagging communities (as defined in the previous section) so that we can enable community-based exploration of the social web. We propose a probabilistic generative model that aims to model the social tagging process by modeling users' activity in the tagging process and their tagging choices to capture community-wide interests.

3.1 Preliminaries

Discovering communities in large linked networks is a rich area. Example community detection approaches include node clustering in large networks [1, 14, 2, 38, 13, 10], web community discovery via content and hyperlink analysis [23, 32, 45, 17], among many others. In this paper, we propose a community-based model that stems from previous works on latent topic models like Latent Semantic Analysis (LSA) [12], Probabilistic Latent Semantic Analysis (pLSA) [20], and Latent Dirichlet Allocation (LDA) [5]. A text-based topic model typically views the words in a text document as belonging to hidden (or "latent") conceptual topics. In this way, a text document is "generated" by an author who samples words from the underlying conceptual topics (e.g., by selecting words from a "football" topic and words from a "finance" topic to write a document about NFL player position vs. salary).

A text-based topic model can be easily adapted to social tagging by considering the document unit to be the collection of all tags applied to a particular resource. We call this collection of tags S_i applied to a resource its *social tagging document*.

[Definition] Social Tagging Document: *For a resource $r \in \mathcal{U}$, we refer to the collection of tags assigned to the resource as the resource's social tagging document S, where S is modeled by the set of tags assigned to the resource: $S = \{tag_j\}$.*

For concreteness, we consider in this paper an adaptation of the LDA model for tag-based modeling (which we shall refer to as TagLDA for clarity).

TagLDA: As in LDA, TagLDA assumes a social tagging document to be generated from a mixture of latent topics, where each topic has a multinomial distribution over the tag vocabulary. Formally, let Φ be a $K \times T$ matrix representing topics, where each ϕ_k is a distribution over tags for topic k, K is the number of topics, and T is the size of tag vocabulary. Similarly, resources are represented by $R \times K$ matrix Θ, where each θ_S is a distribution over topics for social tagging document S.

The TagLDA generative process is as follows:

1. for each topic $z = 1, ..., K$
   - select T dimensional $\phi_z \sim$ Dirichlet(β)
2. for each social tagging document S_i, $i = 1, ..., R$
   - select K dimensional $\theta_i \sim$ Dirichlet(β)
   - For each tag t_j, $j = 1, ..., N_i$
     - Select a topic $z_j \sim$ multinomial(θ_i)
     - Select a tag $t_j \sim$ multinomial(ϕ_{z_j})

A number of approaches have been proposed in the literature to estimate the posterior distributions (the distribution over tags ϕ_k in each topic k and the distribution over topics θ_i for each document) from this generative process such as expectation maximization [5], expectation propagation [30], and Gibbs sampling [18].

3.2 Community-Based Tagging Model

TagLDA provides a foundation for discovering communities in social tags. Fundamentally, however, a social tagging document is a collaborative effort among many taggers, whereas TagLDA is a topic model with no notion of authorship or community. In essence TagLDA can be used to discover social tagging topics over tags, but not social tagging communities over users and their tags, since users are not explicitly modeled in the generation process. Recent work on author-topic models [28] has added the concept of "author" to the LDA model, but fundamentally these models are designed to model text documents that have a single (or a few) authors. In contrast, a social tagging document is the product of (potentially) hundreds of authors. These observations suggest a new approach.

Instead of treating the tag generation process as if tags are generated regardless of user, we propose to couple the tag generation process with that of the user. Concretely, we propose the Community-Based Tagging (CTAG) model for discovering user communities in addition to groupings of tags. Formally, the CTAG model assumes

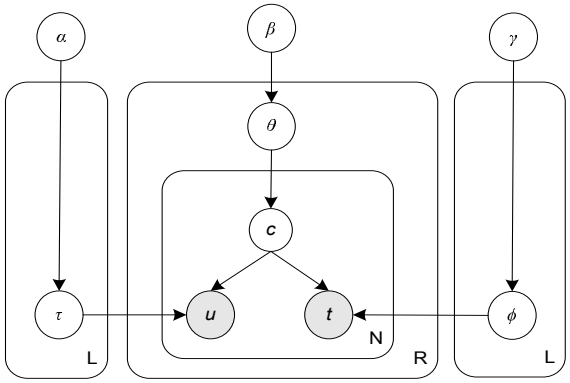


Figure 3: Community-Based Tagging Model

a corpus of R social tagging documents drawn from a vocabulary of T tags and U users, where each social tagging document S_i is of variable length N_i and is composed of both users *and* tags (unlike in the previous definition). As a result, we now redefine the social tagging document to conform to the CTAG modeling approach.

[Definition] Social Tagging Document: *For a resource $r \in \mathcal{U}$, we refer to the collection of tags assigned to the resource as the resource's social tagging document S, where S is modeled by the set of users and the tags they assigned to the resource: $S = \{\langle user_j, tag_j \rangle\}$.*

The CTAG model assumes that the $\langle user, tag \rangle$ pairs in a social tagging document are generated from a mixture of L distinct communities, where each community is a mixture of users with a common world view represented by a mixture of tags. Therefore, the tagging process involves two steps: (i) the selection of a community from which to draw users and tags; and (ii) the selection of the $\langle user, tag \rangle$ pair representing the user influence in the selected community and the preference over tags based on the resource's content, and the tagger's perception of the content.

Let $\mathbf{S}_i$ and $\mathbf{c}$ be vectors of length N_i representing $\langle user, tag \rangle$ pair, and community assignments, respectively, in a social tagging document. The CTAG model generation process is illustrated in Figure 3 and described here:

1. for each community $c = 1, ..., L$
   - Select U dimensional $\tau_c \sim$ Dirichlet(α)
   - Select T dimensional $\phi_c \sim$ Dirichlet(γ)
2. for each social tagging document $\mathbf{S}_i$, $i = 1, ..., R$
   - Select L dimensional $\theta \sim$ Dirichlet(β)
   - For each position $S_{i,j}$, $j = 1, ..., N_i$
     - Select a community $\mathbf{c}_{i,j} \sim$ multinomial(θ_i)
     - Select a user $\mathbf{S}^u_{i,j} \sim$ multinomial($\tau_{\mathbf{c}_{i,j}}$)
     - Select a tag $\mathbf{S}^t_{i,j} \sim$ multinomial($\phi_{\mathbf{c}_{i,j}}$)

The community interest in a resource r_i is represented by the social tagging document's community distribution $\theta_i = \{\theta_{i,j}\}_{j=1}^{L}$ and is sampled from a Dirichlet distribution with parameter $\beta = \{\beta_i\}_{i=1}^{L}$. Each community's world view/ interest is captured by the community tag distribution $\phi_c = \{\phi_{c,i}\}_{i=1}^{|T|}$ and is sampled from a Dirichlet distribution with parameter $\gamma = \{\gamma_i\}_{i=1}^{|T|}$. Finally, each community's user grouping is captured through the community's user distribution $\tau_c = \{\tau_{c,i}\}_{i=1}^{|U|}$ and is sampled from a Dirichlet distribution with parameter $\alpha = \{\alpha_i\}_{i=1}^{|U|}$. The process of generating a social tagging document S_i for a resource r_i fixes the number of tagging actions N_i and for each position $\mathbf{S}_{i,j}$ it samples a community $\mathbf{c}_{i,j}$ from a multinomial distribution with parameter θ_i. This community assignment is then used to draw the $\langle user, tag \rangle$ pair for that position. A user is selected from a multinomial distribution with parameter $\tau_{\mathbf{c}_{i,j}}$ and an associated tag is sampled from a multinomial distribution with parameter $\phi_{\mathbf{c}_{i,j}}$.

Based on the model we can write the likelihood that a position $\mathbf{S}_{i,j}$ is assigned a specific $\langle user, tag \rangle$ pair $\{u, t\}$ as:

$$p(\mathbf{S}_{i,j} = \{u,t\}|\theta_i, \phi, \tau) = \sum_{l=1}^{L} p(\mathbf{S}^u_{i,j} = u|\tau_l) p(\mathbf{S}^t_{i,j} = t|\phi_l) \\ p(\mathbf{c}_{i,j} = l|\theta_i)$$

Furthermore, the likelihood of the complete social tagging document $\mathbf{S}_i$ is the joint distribution of all its variables (observed and hidden):

$$p(\mathbf{S}_i, \mathbf{c}_i, \theta_i, \phi, \tau|\alpha, \beta, \gamma) = \prod_{j=1}^{N_i} p(\mathbf{S}^t_{i,j}|\phi_{\mathbf{c}_{i,j}}) \\ p(\mathbf{S}^u_{i,j}|\tau_{\mathbf{c}_{i,j}}) p(\mathbf{c}_{i,j}|\theta_i)$$

Integrating out the distributions θ, τ and ϕ and summing over $\mathbf{c}_i$ gives the marginal distribution of $\mathbf{S}_i$ given the priors:

$$p(\mathbf{S}_i|\alpha, \beta, \gamma) = \int\int\int p(\theta_i|\beta) p(\phi|\gamma) p(\tau|\alpha) \\ \prod_{j=1}^{N_i} \sum_{\mathbf{c}_{i,j}} p(\mathbf{S}^u_{i,j}|\tau_{\mathbf{c}_{i,j}}) p(\mathbf{S}^t_{i,j}|\phi_{\mathbf{c}_{i,j}}) p(\mathbf{c}_{i,j}|\theta_i) \\ d\phi d\tau d\theta_i$$

Finally our universe of discourse $\mathcal{U}$ consisting of all R social tagging documents occurs with likelihood:

$$p(\mathcal{U}|\alpha, \beta, \gamma) = \prod_{i=1}^{R} p(\mathbf{S}_i|\alpha, \beta, \gamma)$$

3.3 Parameter estimation and inference

The CTAG model provides a generative approach for describing how social tagging documents are constructed. But our challenge is to work in the reverse direction – taking a set of social tagging documents and inferring the underlying model (hidden community distributions). This entails learning model parameters τ, θ, and ϕ (the distributions over communities, users, and tags, respectively).

Although exact computation of these parameters is intractable, several approximation methods have been proposed in the literature for solving similar parameter estimation problems (like in LDA). In this paper, we adopt Gibbs Sampling (see [18] for a thorough treatment) which is a special case of Markov-chain Monte Carlo methods that estimates a posterior distribution of a high-dimensional probability distribution. The sampler draws from a joint distribution $p(x_1, x_2, ..., x_n)$ assuming the conditionals $p(x_i|x_{-i})$ are known, where $x_{-i} = (x_1, ..., x_{i-1}, x_{i+1}, ..., x_n)$.

Let $\mathbf{S}$ and $\mathbf{c}$ be vectors of length $\sum_i^R N_i$ representing $\langle user, tag \rangle$ pair, and community assignments, respectively, for the entire corpus. Also let u and t be user and tag variables. Following the approach used in [18] the joint probability distribution of the CTAG

model can be factored as:

$$p(\mathbf{S}^u, \mathbf{S}^t, \mathbf{c}|\alpha, \beta, \gamma) \quad = \quad p(\mathbf{S}^u|\mathbf{c}, \alpha)p(\mathbf{S}^t|\mathbf{c}, \gamma)p(\mathbf{c}|\beta).$$

We derive the Gibbs sampler's update equation (details not shown for space considerations) for the hidden variables from the joint distribution and arrive at:

$$p(\mathbf{c}_i = l|\mathbf{c}_{\neg i}, \mathbf{S}^t, \mathbf{S}^u) \quad \propto \quad \frac{n^u_{l,\neg i} + \alpha_u}{\sum_{u=1}^{U} n^u_{l,\neg i} + \alpha_u} \times \frac{n^t_{l,\neg i} + \gamma_t}{\sum_{t=1}^{T} n^t_{l,\neg i} + \gamma_t} \times \frac{n^l_{S,\neg i} + \beta_l}{\left(\sum_{l=1}^{L} n^l_S + \beta_l\right) - 1} \tag{1}$$

where $n^{(\cdot)}_{(\cdot),\neg i}$ is a count excluding the current position assignments of $\mathbf{c}_i$ (e.g., $n^t_{l,\neg i}$ is the count of tag t generated by the l-th community excluding the current position).

4. COMMUNITY-BASED RANKING

In the previous section we presented the CTAG model for discovering implicit social tagging communities. We now turn our attention to leveraging this information for community-based exploration of socially tagged documents. Our goal is to leverage the discovered community structure to implicitly connect users, tags, and resources for more effective information exploration and discovery.

After applying the CTAG model to a collection of user, tag, resource tuples we have as output several distributions:

- For each community, we have a probability distribution over all users $\tau_c = \{\tau_{c,i}\}_{i=1}^{|U|}$
- For each community, we have a probability distribution over all tags $\phi_c = \{\phi_{c,i}\}_{i=1}^{|T|}$
- For each resource, we have a probability distribution over communities $\theta_i = \{\theta_{i,j}\}_{j=1}^{L}$,

Based on these discovered distributions, we can, for example, identify implicitly related users and implicitly related tags based on their common community membership. These relationships can be used to automatically suggest related tags, to recommend unknown users (and their collection of bookmarks) to interested users, and so forth. Similarly, we can identify implicitly related resources based on their community distribution (e.g., to identify similar resources based on the communities that are interested in them), and support other forms of social exploration.

While the possibilities are quite large for applying the discovered community-based information from the CTAG model, we examine in the rest of this section two approaches for ranking resources based on the community's perspective. Concretely, we consider: (i) a query-community ranking approach that maps a user's topical interest (expressed as a query) to resources preferred by communities with a similar topical interest; and (ii) a user-community ranking approach that re-ranks all resources based on the user's implicit community, regardless of the query. In both cases, we are interested to examine if the discovered implicit community structure can enable more effective ranking than traditional (non-community) based approaches.

4.1 Query-Community Ranking

In the first approach, we aim to boost a baseline resource ranking by two factors: (i) the query term importance in each community; and (ii) the resource importance in each community. The first factor boosts the ranking of resources that contain query terms considered important by the community regardless of document's community preference. The second factor boosts the ranking of resources that are preferred by the community regardless of them containing the query term. The net effect boosts the ranking of resources that are both preferred by the community and contain query terms that are considered important by the community.

This query-community ranking is defined as product of likelihoods of query terms relevance to resources as follows:

$$Score(S, Q) = \prod_{t \in Q} p(S|t)$$

Now, $p(S|t)$, the resource relevance to a query is computed over all communities using the two factors mentioned above, the community preference for the document and the community preference for the tag as follows:

$$p(S|t) = \sum_{c=1}^{L} p(S|c)p(c|t) \tag{2}$$

Using Bayes' rule we further expand each factor to be:

$$p(S|c) = \frac{p(c|S)p(S)}{p(c)} \quad \text{and} \quad p(c|t) = \frac{p(t|c)p(c)}{p(t)}$$

and by substituting into (2) we finally get:

$$p(S|t) = \sum_{c=1}^{L} \frac{p(c|S)p(t|c)p(S)}{p(t)}$$

The quantities $p(c|S)$ and $p(t|c)$ are readily available from the CTAG model results:

$$p(c|S) = \theta^c_S \quad \text{and} \quad p(t|c) = \phi^t_c$$

The document prior probability $P(S)$ and the tag prior probability $P(t)$ are collection dependent and we compute them as follows:

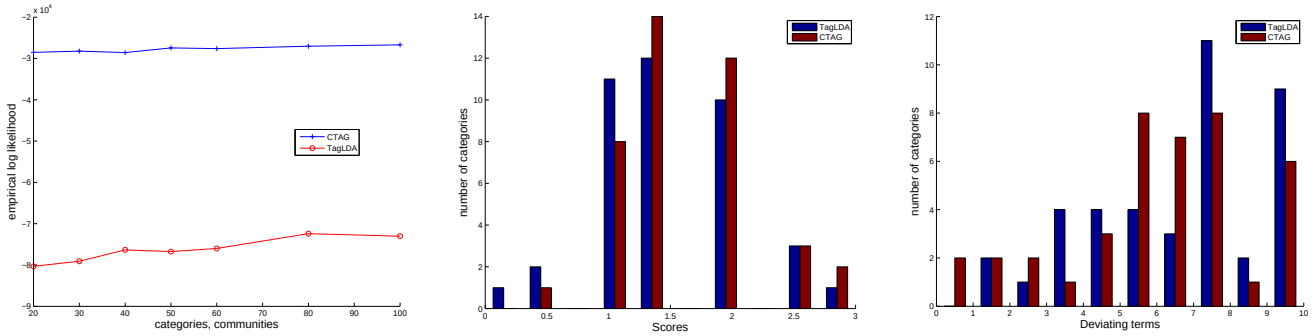
$$P(S) = \frac{|S|}{\sum_{i \in R} |S_i|} \quad \text{and} \quad P(t) = \frac{tf(t)}{\sum_{i \in T} tf(i)}$$

where, $|S|$, is the length of document S, and $tf(t)$ is the count of tag t.

4.2 User-Community Ranking

In the second approach, we aim to boost resource ranking by user community information. To that end, we consider the community membership for each user as determined by our model. Knowing a user's community strength, we can favor resources that are most preferred from the user's community, even if the user has never tagged the resource. This approach constitutes two factors: one that accounts for community preference for a resource and the second accounts for user membership in that community. This user-community ranking is defined as follows:

$$Score(S, u) = \sum_{c=1}^{L} p(S|c)p(c|u)$$



(a) Empirical likelihood (b) Community tag coherence scores (c) Deviating terms

Figure 4: Comparing CTAG to TagLDA

Using Bayes' rule again we expand $p(S|c)$ and $p(c|u)$ into:

$$p(c|u) = \frac{p(u|c)p(c)}{p(u)} \quad \text{and} \quad p(S|c) = \frac{p(c|S)p(S)}{p(c)}$$

Then substituting into the previous equation we get:

$$Score(S,u) = \sum_{c=1}^{L} \frac{p(u|c)p(c|S)p(S)}{p(u)}$$

These quantities are, again, readily available from the CTAG model results:

$$p(u|c) = \tau_c^u \quad \text{and} \quad p(c|S) = \theta_S^c$$

The document prior probability is computed is in the previous section and the user prior probability is computed as:

$$P(u) = \frac{uf(u)}{\sum_{i \in U} uf(i)}$$

where $uf(u)$ is the count of user u in the collection.

4.3 Rank Aggregation

In both the case of the query-community score and the user-community score, we can combine each individual score with a baseline query-resource score to arrive at a final score for each social tagging document with respect to a query. In this paper, as a baseline ranking approach, we adapt the popular BM25 retrieval model to the context of retrieval on social bookmarking systems [36]. For a user who is interested in searching the web of socially tagged resources, we can adapt the BM25 ranking over $\mathcal{U}$ for a query Q by scoring each social tagging document S:

$$Score_{BM25}(S,Q) = \sum_{t \in Q} IDF(t) \frac{f(t,S)(k_1+1)}{f(t,S) + k_1(1 - b + b\frac{|S|}{avgL})}$$

$$IDF(t) = \log \frac{R - n(t) + 0.5}{n(t) + 0.5}$$

where, $f(t,S)$ = frequency of tag t in social tagging document S, $|S|$ = total tags in S, $avgL$ = average length of documents in $\mathcal{U}$, R = total number of documents, $n(t)$ = number of documents containing tag t. $k1$ and b are free parameters which we take to be their typical settings: $k1 = 2$, $b = 0.75$. BM25 provides a baseline for ranking resources by considering the presence of query terms in the social tagging document, but makes no attempt to incorporate community or latent topic information.

To combine this baseline with the query-community score and the user-community score, we rely on rank aggregation, which is the task of combining voters' rankings of a set of candidates to obtain a single ranking for the set. It is a well known problem encountered in many contexts; especially in social choice theory. In our case the candidates are social web documents and the voters are the BM25 and the proposed community-based ranking functions. To combine the rankings produced by the ranking functions, we adopt a simple positional method known as Borda's Rule [11]. In Borda's rule each candidate is awarded a point for each competitor candidate ranked below it. Candidates are finally ranked by their accumulated points.

5. EXPERIMENTS

In this section, we evaluate the quality of the Community-Based Tagging Model (CTAG) model and the community-based ranking approach over the social bookmarking system Delicious. We first investigate the quality of the discovered communities using CTAG by comparing each community's tag distribution to tag distributions derived from the non-user based TagLDA approach. Next, we evaluate the proposed community-based ranking to three state-of-the-art retrieval models: (i) BM25; (ii) Cluster-based retrieval using K-means clustering; and (iii) LDA-based retrieval.

For the dataset, we crawled Delicious starting from a set of popular tags. The crawler has discovered 607,904 unique tags, 266,585 unique Web pages annotated by Delicious, and 1,068,198 unique users. Of the 266,585 total Web pages, we have retrieved the full HTML for 47,852 pages. We filter this set to keep only pages in English with a minimum length of 20 words, leaving us with 27,572 Web pages with 16,216 unique tags. We preprocessed both corpora by removing rare tags and users (with only a single occurrence), and stemming the tags.

5.1 Model and Clusters Quality

In our first set of experiments, we compare the community-based tagging model (CTAG) against the tag-based version of Latent Dirichlet Allocation (TagLDA). Recall that TagLDA discovers latent topics of related tags whereas CTAG simultaneously discovers both communities of related users *and* the tags that each community is most interested in (probabilistically). Since TagLDA has no notion of users or user-based communities, our goal here is to first

Table 2: Delicious: Communities, their top tags, and their top users

	top tags	top users and their top tags
Comm 0	dictionari translat encyclopedia slang thesauru grammar spanish french acronym linguist etymolog answer diccionario vocabulari ital	user 11985: web,freestylelanguag,video,onlin,italian,german,dictionari, translat... user 73941: word, cultur, resourc, languag, internet, vocabulari, buzzword, lexicon... user 24880: photographi, fileshar, folksonomi, english, copyright,resourc, document... user 8846: gener, usabl, english, audio, articl, publish, write, literaci, dictionari, ...
Comm 1	mindmap diagram brainstorm visio oreilli shoe whiteboard anatomi flowchart mind uml bodi brain conceptmap wirefram meet graphicorga	user 8226: podcast,visual,audio,music,draw, brainstorm,folksonomi,talotool... user 36411: document, new, innov, photo, creativ, idea, visual, health, mindmap... user 30704:mindmap, grid, video, imag, art, technolog, anim, elearn, portfolio... user 12696: photo, wiki, audio, mindmap, collabor, brainstorm, sound, imag...
Comm 2	dn bank financi credit bill budget auction ebai date domain loan lend invoic palett weather calcul trade currenc clock frugal	user 7587: innov,energi, infrastructur, usa, blog, nielsen, communicat, media... user 18375:program, articl, cool, bank, howto, product, onlin, daili,blog, bill .. user 65615: develop, softwar, program, job, search, callcent, book, servic... user 2165: interest, program, list, websit, review, financ, altern, lifestream, admin...
Comm 3	distro vmware debian tweak emul recoveri driver vm vista laptop kernel wine registri livecd gentoo boot sandbox usb thunderbird uninstal	user 55702:refer, freeservic, search, orpdw, installed, educ, applic, secur, sourceforg... user 4959: tip, develop, virtual, architectur, hack, share, technolog, host, widget... user 24559:lifehack, todo, develop, technolog, applic, tutori, util, network, audio... user 23732: secur, extens, tutori, howto, educ, develop, linux, ubuntu, hardwar...
Comm 4	airlin flight seat airfar airplan deal ticket coupon bargain cheap hotel question vacat trip aviat discount sport fly metafilt price	user 2147: busi, network, televis, book, guid, directori, mashup, rate, locat, mobil.. user 59837: onlin, web, homeschool, recip, game, teach, travel, food, hotel... user 65681:travel, airlin, blog, transport, flight, ticket, airfar, vacat, refer, map, deal... user 5187: map, airlin, blog, flight, refer, web20, aggreg, tool, resourc, internet, social,...

understand how the models compare with respect to the discovered groups of related tags.

Setup: For both TagLDA and CTAG, we modified a popular public implementation of LDA distributed in the Mallet toolkit [29]. For TagLDA, we set the model hyperparameters to the default values in the Mallet toolkit ($\alpha = 50/K, \beta = 0.01$) with optimization enabled. For CTAG, we experimented with several combinations of hyperparameters optimized by the fixed-point iteration method in [31]. The results we compare with other models are run with hyperparameters ($\alpha = 0.1, \beta = 1, \gamma = 0.1$) and optimization enabled. For both TagLDA and CTAG, a Gibbs sampler starts with randomly assigned communities (or topics for TagLDA), runs for 2000 iterations with optimization every 50 iterations and an initial burnin-period of 250 iterations. We vary the number of communities from 20 to 100.

Empirical Likelihood of Unseen Data: We evaluate each model's generalization to unseen data using the empirical likelihood method [25]. To compute empirical likelihood, we generate 1000 documents based on each model's generative process. We then build a multinomial over the vocabulary space from these samples. Finally we compute the empirical likelihood of a held-out testing set using the obtained multinomials over the vocabulary space. A model that results in a higher empirical likelihood is essentially a "better fit" to the observed data.

We plot the empirical likelihood results in Figure 4 for both TagLDA and CTAG. The y-axis shows the empirical log likelihood and the x-axis show the number of communities (or topics for TagLDA). The CTAG model performs better than TagLDA. It peaks around 50 communities, then decreases slightly and stabilizes. Likewise, the performance of TagLDA peaks around 40 categories, decreases slightly, then peaks again at 80 categories. Based on these results, we show that the CTAG model is better suited to modeling social tagging data than the LDA model through better handling of unseen data.

User Study of Discovered Communities: We further conduct a user study to judge the coherence of the tag-based communities uncovered by the models. Communities from each model were anonymized and put in random order and presented to four human judges. Each judge is asked to determine the coherence from each group of tags by trying to detect a theme from its top 10 tags. Coherence is graded on a $0 - 3$ scale with 0 being poor coherence and 3 excellent coherence. The judges are also asked to report the number of terms that deviate from the theme they thought the community represented. The results of this user study are shown in Figure 4(b) and Figure 4(c). Figure 4(b) shows the number of tag-based communities from each model and the coherence scores they received. Notice that CTAG has higher number of categories receiving a score of 2 or higher compared to TagLDA. We can also see that CTAG has lower number of categories receiving a score of 1 or lower compared to TagLDA. Figure 4(c) shows the number of categories from each model versus the number of deviating terms. Again CTAG has a higher number of categories containing small number of deviating terms compared to TagLDA and a lower number of categories containing large number of deviating terms. Based on this evaluation, we see that the CTAG model is better suited to modeling social tagging data than the LDA model by discovering more coherent collections of tags.

Example Communities Discovered: To illustrate we show a sample of communities, their top tags, and their top users along with their associated tags in Table 2.

5.2 Ranking Over Socially Tagged Resources

In the second set of experiments, we turn to the challenge of testing if the discovered communities can enhance the exploration of the social web through the community-based ranking models introduced in Section 4. To compare the community-based ranking model, we consider two alternative state-of-the-art retrieval models: (i) Cluster-based retrieval using K-means clustering; and (ii) LDA-based retrieval. While these retrieval models have been developed in the context of text-based retrieval, we adapt each to the context of retrieval on social bookmarking systems as described in the following brief sections.

Cluster-Based Retrieval (K-means): The first approach is a tag-based implementation of cluster-based retrieval introduced by Liu and Croft [27]. Cluster-based retrieval hypothesizes that by grouping text documents (in our case, social tagging documents), the quality of ranking can be improved by smoothing each document with the rest of the documents in the cluster (in essence, asserting that similar documents will satisfy the same information need). In practice, we use K-means clustering to cluster all social tagging documents; we set $k = 40$ based on the results of the previous experiment. Documents in each cluster are then combined to build a unigram language model, i.e., a multinomial distribution over its

vocabulary space. Ranking in this case is based on clusters instead of documents.

$$Score(cluster, Q) = \prod_{t \in Q} p(t|cluster).$$

The quantity $p(t|cluster)$ is computed from a cluster language model smoothed by a background model as follows:

$$p(t|cluster) = \lambda \frac{tf(t, cluster)}{\sum_t tf(t, cluster)} + (1 - \lambda) \frac{tf(t, Coll)}{\sum_t tf(t, Coll)}$$

where $tf(t, cluster)$ is the count of tag t in the cluster and $tf(t, Coll)$ is the count of tag t in the entire collection. The free parameter ($\lambda = 0.5$) controls the smoothing proportion. The cluster-based ranking can then be combined with the per-document BM25 score using rank aggregation as in the case of community-based ranking.

LDA-Based Retrieval: The second approach we consider follows Wei and Croft [39] to incorporate the LDA based document representation for retrieval. Given the inferred distributions from TagLDA, we can define an LDA-based ranking function as follows:

$$Score(S, Q) = \prod_{t \in Q} p(t|S)$$

Now, $p(t|S)$, the query likelihood given the document is computed over all tag topics using two factors: the document preference for the topic and the topic preference for the tag as follows:

$$p(t|S) = \sum_{z=1}^{K} p(z|S)p(t|z) \quad (3)$$

The quantities $p(z|S)$ and $p(t|z)$ are available from the TagLDA model results, $p(z|S) = \theta_S^z$ and $p(t|z) = \phi_z^t$. The LDA-based scores can then be combined with the per-document BM25 scores using rank aggregation.

5.2.1 Tag-based Retrieval

To evaluate the quality of community-based ranking and to be fair across all models, we first consider retrieval using only tags (since BM25, LDA, and K-means do not model the user as CTAG does). We select three sets of tags with the following criteria:

Rare tags: Six rare tags, that is tags that occur on at most 5 resources.

Unambiguous tags: Eight pairs of unambiguous tags, where we pair the tag *"tool"* with a number of tags such as *"finance", "music", "health", "social", "game"*, making the pair, *"music tool"*, very specific in what is expected to be retrieved.

Popular tags: Twelve popular tags, tags like *"new", "program", "resource", "howto", "blog"*.

For each of these tag sets we retrieve the top ten relevant documents per query using each ranking model – BM25, BM25+Kmeans, BM25+LDA, and BM25+CTAG. The results for each query are presented to human judges to determine their relevance to the query on a scale of 1 to 5 with 1 being least relevant and 5 most relevant. The judgements scores are analyzed using Normalized Discounted Cumulative Gain (NDCG)[21]. For a list of graded resources, NDCG computes the gain of each resource in the list based on its grade and rank and accumulates the gains over the list up to a specified position. Table 3 presents the NDCG@10 for each

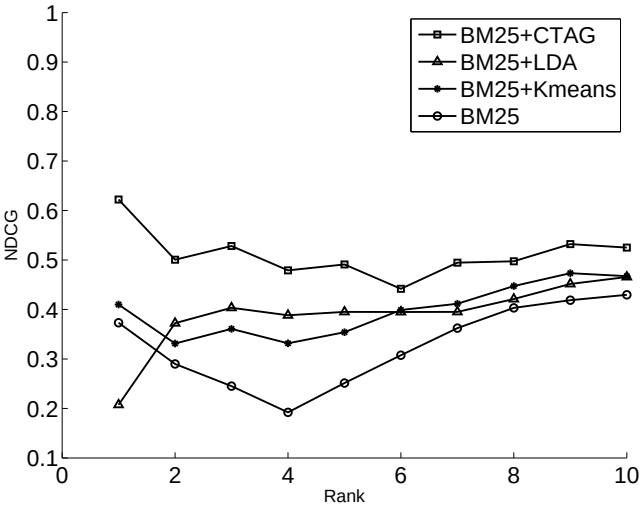


Figure 5: Ranking Quality for Unambiguous Queries

ranking model across the three types of queries, as well as the percent change by the proposed community-based ranking model BM25+CTAG versus the other three approaches.

First, consider the set of rare tags. Suppose our collection has 3 documents that carry the tag x. When a user searches for this tag using a traditional retrieval methods, e.g. BM25, those 3 documents will be returned as relevant and all other documents are given a score of zero, or a corpus wide smoothing score. However, there might be documents in the collection that do not carry the query tag but are relevant to the query, e.g., documents tagged with synonyms of the query term, or misspellings, or topically relevant tags. Community-based tag grouping (as in CTAG) could help improve results for this kind of query and our results support this conjecture. As Table 3 shows, the CTAG model results in the best ranking quality for rare tag queries, improving on BM25 by 20%, improving on K-means by 4%, and on LDA by 7%. We attribute this improvement to the ability of the CTAG model to better fit social tagging data than LDA or K-means. See Figure 6 for more detailed NDCG results for rank positions up to 10.

Second, for the set of unambiguous tags, the intuition is that a tag such as "tool" is popular, general and belongs uniformly to many communities while the the other tags are specific and could be prominent in small number of communities. For a traditional retrieval model the results are dominated by documents relevant to the general term due to high term frequency in document that might not be relevant to the second term. At the same time, the scores of the more specific term are not prevalent enough to make it to top positions in the retrieved list. Community-based scores can help bring those documents that are considered valuable to the second terms's community up in the list. As Table 3 shows, the CTAG model results in the best ranking quality for unambiguous tag queries, improving on BM25 by 22%, improving on K-means by 12%, and on LDA by 12%. See Figure 5 for more detailed NDCG results for rank positions up to 10.

Finally, for the set of popular tags, their popularity makes them belong uniformly to many communities making the community structure, in this case, of little benefit. The community structure might actually degrade the retrieval performance by promoting documents that are too general and perceived uniformly across communities, which we suspect to be the case in our results. Another

Table 3: Tag-Based retrieval results

	NDCG@10				%change by BM25+CTAG over		
	BM25	BM25+Kmeans	BM25+LDA	BM25+CTAG	BM25	BM25+Kmeans	BM25+LDA
Unambiguous tags	0.42	0.46	0.46	0.52	0.22	0.12	0.12
Rare tags	0.67	0.78	0.76	0.81	0.2	0.04	0.07
Popular tags	0.65	0.67	0.69	0.58	-0.09	-0.12	-0.15

issue with this kind of queries, is the difficulty to evaluate relevance when query terms are vague. This was evident in the disagreements among judges scores for popular tag results.

To test judge biases in scoring the results for the different queries we use the Wilcoxon signed rank test [40], which given two paired samples of measurements, tests if the differences come from a symmetric distribution with zero median against the alternative that differences do not have a zero median. In our case if the differences have a zero median, we can conclude that the judges biases are not significant and that there is significant agreement on how the results are ranked. The Wilcoxon signed rank test results are shown in Table 4. When the P-value < 0.05, we reject the hypothesis of zero median and conclude that there is significant disagreement among judges scores. This is seen only in the case of popular tags.

Table 4: Wilcoxon signed rank test of agreement between judges scores with 95% confidence

Query group	P-value(two-tailed)
Unambiguous tags	0.10
Rare tags	0.39
Popular tags	< 0.0001
User-based retrieval tags	0.39

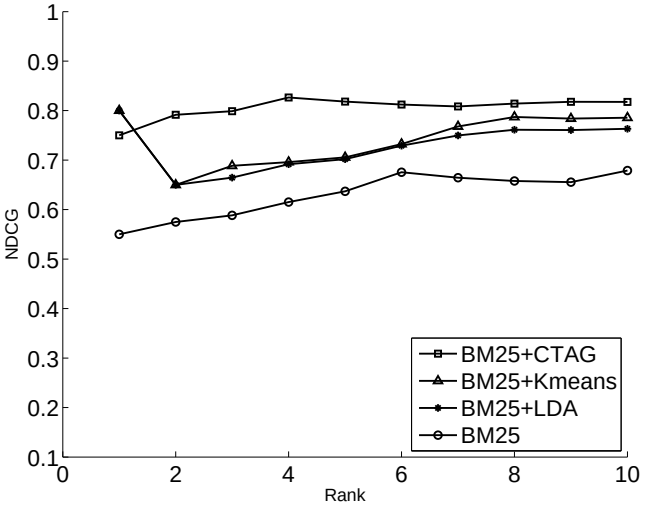


Figure 6: Ranking Quality for Rare Tag Queries

5.2.2 User-based Retrieval

Now that we have seen how community-based ranking can improve tag-based retrieval, we next consider how user-based retrieval can be improved. The goal of this section is to show the benefit of user modeling in social tagging as is done in the CTAG model.

To that end, we select five users that exhibit interest for some of the tags we used in the previous experiment, *i.e.* top users from the most representative community for the tag. For each user and tag combination, we retrieve the top ten relevant documents. These results are judged for relevance as was previously done on a scale of 1 to 5. The NDCG of judges results are as shown in Table (5). The results for CTAG model include both the query-community (CTAG) ranking and the user-community (CTAG(U)) ranking. Notice that CTAG performs best with 47%, 43% and 60% improvement over BM25, K-means, and LDA respectively. These improvements show that user-based community structure uncovered by the CTAG model helps improve ranking of tagged resources.

6. CONCLUSIONS

In this paper, we have proposed and evaluated a generative community based probabilistic tagging model in which users belong to implicit groups of interest (e.g., students, sports fans) and probabilistically select tags with which to bookmark resources. Coupled with Gibbs sampling parameter estimation, the community-based model can automatically uncover these communities of implicitly related users and their associated tags. We have seen how the community-based model improves the empirical likelihood of held-out test data and discovers more coherent interest-based communities compared to LDA. We have also developed a novel community-based ranking model for effective community-based exploration of socially-tagged Web resources. Compared to BM25, Cluster-based retrieval using K-means clustering, LDA-based retrieval, we find that the proposed ranking model results in a significant improvement over these alternatives (from 7% to 22%) in the quality of retrieved pages.

Although our approach suggests an important role for socially contributed data in advancing information discovery, there are a number of limitations to its application and generalization to social tagging systems at large. First, LDA based approaches, in general, including our CTAG model require global knowledge and perform many iterations to uncover latent variables. Hence, using them on-line is difficult. Second, our CTAG model, as does LDA, assumes a fixed number of latent variables and does not consider the temporal aspect of tagging. Therefore, it cannot capture growth and evolution. Third, our assumption of global user communities does not capture individual user behavior. In addition, the lack of standard corpora for social tagging data makes evaluating and comparing results of different research methods difficult. And, results based on human judges of individually collected corpora need to be verified for generalization to different social tagging systems.

However, some of these limitations can be overcome. A combination of a both on-line and off-line approach can solve the processing requirements of LDA-based models. Also, there are methods for dynamically discovering the number of latent variables (see for example [4]). As future work, we will incorporate time into the CTAG model to capture community evolution over time. We will also investigate ways to construct individual user models in addition to the global user communities.

Table 5: User-based retrieval results

NDCG@10				%change by BM25+CTAG(U) over		
BM25	BM25+Kmeans	BM25+LDA	BM25+CTAG(U)	BM25	BM25+Kmeans	BM25+LDA
0.39	0.40	0.36	0.58	0.47	0.43	0.60

7. ACKNOWLEDGMENTS

The first author is supported by a scholarship from the Ministry of Manpower, Oman. This work is partially supported by faculty startup funds from Texas A&M University and the Texas Engineering Experiment Station.

8. REFERENCES

[1] R. Albert and A. Barabási. Statistical mechanics of complex networks. *Reviews of Modern Physics*, 74(1):47–97, 2002.

[2] R. Albert, H. Jeong, and A.-L. Barabasi. The diameter of the world wide web. *Nature*, 401:130, 1999.

[3] S. Bao, G. Xue, X. Wu, Y. Yu, B. Fei, and Z. Su. Optimizing web search using social annotations. In *WWW*, 2007.

[4] D. M. Blei, T. L. Griffiths, M. I. Jordan, and J. B. Tenenbaum. Hierarchical topic models and the nested chinese restaurant process. In *NIPS*, 2004.

[5] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent dirichlet allocation. In *JMLR*, volume 3, pages 993–1022, April 2003.

[6] C. H. Brooks and N. Montanez. Improved annotation of the blogosphere via autotagging and hierarchical clustering. In *WWW*, 2006.

[7] C. Cattuto, A. Baldassarri, V. D. P. Servedio, and V. Loreto. Vocabulary growth in collaborative tagging systems, 2007.

[8] C. Cattuto, A. Baldassarri, V. D. P. Servedio, and V. Loreto. Emergent community structure in social tagging systems. *CoRR*, 2008.

[9] C. Cattuto, V. Loreto, and L. Pietronero. Collaborative tagging and semiotic dynamics, 2006.

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

[11] D. Coppersmith, L. Fleischer, and A. Rudra. Ordering by weighted number of wins gives a good ranking for weighted tournaments. In *SODA*, pages 776–782, 2006.

[12] S. C. Deerwester, S. T. Dumais, T. K. Landauer, G. W. Furnas, and R. A. Harshman. Indexing by latent semantic analysis. *JASIST*, 41(6):391–407, 1990.

[13] M. Faloutsos, P. Faloutsos, and C. Faloutsos. On power-law relationships of the internet topology. In *SIGCOMM*, 1999.

[14] M. Girvan and M. E. J. Newman. Community structure in social and biological networks. *PNAS*, 99:7821, 2002.

[15] S. Golder and B. A. Huberman. The structure of collaborative tagging systems, Aug 2005.

[16] H. Halpin, V. Robu, and H. Shepherd. The complex dynamics of collaborative tagging. In *WWW*, 2007.

[17] T. H. Haveliwala. Topic-sensitive pagerank. In *WWW*, pages 517–526, May 2002.

[18] G. Heinrich. Parameter estimation for text analysis. Technical report, 2004.

[19] P. Heymann, G. Koutrika, and H. Garcia-Molina. Can social bookmarking improve web search? In *WSDM*, 2008.

[20] T. Hofmann. Probabilistic latent semantic indexing. In *SIGIR*, pages 50–57, 1999.

[21] K. Järvelin and J. Kekäläinen. Cumulated gain-based evaluation of ir techniques. *TOIS*, 20(4):422–446, 2002.

[22] J. M. Kleinberg. Authoritative sources in a hyperlinked environment. *J. ACM*, 46(5):604–632, 1999.

[23] H. Li, Z. Nie, W.-C. Lee, L. Giles, and J.-R. Wen. Scalable community discovery on textual data with relations. In *CIKM*, pages 1203–1212, 2008.

[24] R. Li, S. Bao, Y. Yu, B. Fei, and Z. Su. Towards effective browsing of large scale social annotations. In *WWW*, 2007.

[25] W. Li and A. Mccallum. Pachinko allocation: Dag-structured mixture models of topic correlations. In *ICML*, 2006.

[26] X. Li, L. Guo, and Y. E. Zhao. Tag-based social interest discovery. In *WWW*, pages 675–684, 2008.

[27] F. Liu, C. Yu, and W. Meng. Personalized web search for improving retrieval effectiveness. *TKDE*, 16(1):28–40, 2004.

[28] A. Mccallum, X. Wang, and A. Corrada-Emmanuel. Topic and role discovery in social networks. In *IJCAI*, 2005.

[29] A. K. McCallum. Mallet: A machine learning for language toolkit. http://mallet.cs.umass.edu, 2002.

[30] T. Minka and J. Lafferty. Expectation-propagation for the generative aspect model. In *UAI*, pages 352–359, 2003.

[31] T. P. Minka. Estimating a dirichlet distribution. 2003.

[32] L. Nie, B. D. Davison, and B. Wu. From whence does your authority come?: utilizing community relevance in ranking. In *AAAI*, pages 1421–1426, 2007.

[33] L. Page, S. Brin, R. Motwani, and T. Winograd. The pagerank citation ranking: Bringing order to the web, 1999.

[34] A. Plangrasopchok and K. Lerman. Exploiting social annotation for automatic resource discovery. In *IIWeb*, 2007.

[35] D. Ramage, P. Heymann, C. D. Manning, and H. G. Molina. Clustering the tagged web. In *WSDM*, 2009.

[36] S. Robertson and H. Zaragoza. The probabilistic relevance method: Bm25 and beyond. In *SIGIR Tutorial*, 2007.

[37] C. Veres. The language of folksonomies: What tags reveal about user classification. *NLDB*, pages 58–69, 2006.

[38] D. J. Watts and S. H. Strogatz. Collective dynamics of 'small-world' networks. *Nature*, (6684).

[39] X. Wei and W. B. Croft. Lda-based document models for ad-hoc retrieval. In *SIGIR*, 2006.

[40] F. Wilcoxon. Individual comparisons by ranking methods. *Biometrics Bulletin*, 1(6):80–83, 1945.

[41] X. Wu, L. Zhang, and Y. Yu. Exploring social annotations for the semantic web. In *WWW*, 2006.

[42] S. Xu, S. Bao, B. Fei, Z. Su, and Y. Yu. Exploring folksonomy for personalized search. In *SIGIR*, 2008.

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

[44] D. Zhou, J. Bian, S. Zheng, H. Zha, and C. L. Giles. Exploring social annotations for information retrieval. In *WWW*, 2008.

[45] Y. Zhou and J. Davis. Community discovery and analysis in blogspace. In *WWW*, pages 1017–1018, 2006.