World Wide Web (2009) 12:399–420
DOI 10.1007/s11280-009-0068-2

Social Propagation: Boosting Social Annotations for Web Mining

Shenghua Bao · Bohai Yang · Ben Fei · Shengliang Xu · Zhong Su · Yong Yu

Received: 29 November 2008 / Revised: 6 April 2009
Accepted: 13 July 2009 / Published online: 1 August 2009

Abstract This paper is concerned with the problem of boosting social annotations using propagation, which is also called *social propagation*. In particular, we focus on propagating social annotations of web pages (e.g., annotations in Del.icio.us). Social annotations are novel resources and valuable in many web applications, including web search and browsing. Although they are developing fast, social annotations of web pages cover only a small proportion (<0.1%) of the World Wide Web. To alleviate the low coverage of annotations, a general propagation model based on Random Surfer is proposed. Specifically, four steps are included, namely basic propagation, multiple-annotation propagation, multiple-link-type propagation, and constraint-guided propagation. The model is evaluated on a dataset of 40,422 web pages randomly sampled from 100 most popular English sites and ten famous academic sites. Each page's annotations are obtained by querying the history interface of Del.icio.us. Experimental results show that the proposed model is very effective in increasing the coverage of annotations while still preserving novel properties of social annotations. Applications of propagated annotations on web search and classification further verify the effectiveness of the model.

Categories and Subject DescriptorsH. [**Information Systems**]: Miscellaneous; H.3.1 [**Information Storage and Retrieval**]: Content Analysis and Indexing; I.2.6 [**Artificial Intelligence**]: Learning.

General Terms Algorithms, Experimentation, Human Factors.

S. Bao (✉) · B. Yang (✉) · S. Xu (✉) · Y. Yu
Shanghai Jiao Tong University, Shanghai 200240, China
e-mail: shhbao@apex.sjtu.edu.cn
e-mail: yangbohai@apex.sjtu.edu.cn
e-mail: slxu@apex.sjtu.edu.cn

Y. Yu
e-mail: yyu@apex.sjtu.edu.cn

B. Fei · Z. Su
IBM China Research Lab, Beijing 100094, China

B. Fei
e-mail: feiben@cn.ibm.com

Z. Su
e-mail: suzhong@cn.ibm.com

Keywords social annotations · propagation · hyperlink · sitemap-tree link

1 Introduction

Web 2.0 is leading a new revolution on the World Wide Web, where social annotations have become more and more popular. On Web 2.0 sites, like Del.icio.us and Flickr, various kinds of web resources are popularly annotated by web users with freely chosen keywords, which are also known as social annotations or tags. Such annotations not only benefit web user, but also are helpful in many web applications. For example, the social annotations of web pages can be used to improve web search [1, 35, 40], enterprise search [7], personalized search [23], web browsing [18], and semantic web [34, 38]. Annotations given to other resources are also useful, e.g., annotations of Flickr which are assigned to pictures can be effectively utilized for event visualization and detection [8, 27].

While social annotations are useful, many applications of social annotations suffer from the annotation sparseness problem [1, 7, 35]. The web resources with social annotations are still limited on the World Wide Web. Take Del.icio.us as an example, more than 1 million web users collected over 10 million web pages with millions of social annotations.[1] However, compared with 1.173 billion web users[2] and about 30 billion[3] web pages on WWW, the ratio of both social annotators and annotated web pages still remains less than 0.1%. The same annotation sparseness problem also appears in the most popular part of web pages on WWW. To name a few, based on our study (random sample of about 10,000 pages), more than 75% of the pages in ODP are not annotated by web users; about 80% of pages within the top 100 most popular English sites[4] have no annotation at all.

As a result, it is urgent to boost the social annotations and alleviate the annotation sparseness problem. Important as it is, to the best of our knowledge, no previous work has addressed this problem. In the rest of the paper, we confine ourselves to the discussion of social annotations of web pages since they are most representative and valuable part in Web 2.0. Unless otherwise specified, the social annotations refer to those of web pages.

A simple way to boost social annotations is to provide more convenient tools to reduce the cost of assigning social annotations and encourage more users to participate. However, due to the annotation's power-law distribution and stable phenomenon [12, 18], some popular pages may have achieved stable state and still keep receiving new annotations (also known as "rich get richer") while a great amount of the rest pages may only have limited annotations and still suffer from the sparseness problem. Furthermore, due to the habit of web users, some pages rarely get annotated since they can be easily accessed by some other important pages, or they are new pages that are not easy to discover [1].

Another way to boost social annotations is to extract the key terms [10] or generate annotations [5] from the web pages and view them as good summaries. Such methods can be applied to each page on the Web. However, since it has no user knowledge at all, it cannot guarantee that the generated key terms have the same novel property and high quality as social annotations.

To alleviate the annotation sparseness problem effectively, in this paper, we propose a method, called *social propagation*, to boost social annotations automatically using

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

[2] http://www.internetworldstats.com/stats.htm.

[3] http://www.boutell.com/newfaq/misc/sizeofweb.html.

[4] http://www.alexa.com/site/ds/top_sites?ts_mode=lang&lang=en.

propagation. The idea of propagation has been used in other applications, like page quality evaluation [16, 24, 36], relevance estimation [6, 25, 30, 33]. While sharing similar basic idea of propagating via web links, previous models are not suitable for social propagation due to different propagation resource and purpose. When improving the coverage of social annotations, a good social propagation model should preserve the novel properties of social annotations, e.g., keyword property indicating that annotations are good summary of the corresponding web page [1, 7, 18, 35], popularity property indicating that the amount of annotations represent a page's popularity [1, 8, 27], and complexity property which expresses the power-law distribution of social annotations [12, 18].

In particular, we propose a general social propagation model based on the Random Surfer. More specifically, the propagation model consists of four surfing steps:

1) **Basic Surfing**: The basic surfing is used to model the propagation of a single annotation via the same type of links. Unlike previous models of PageRank [24] and Relevance Propagation [30], the basic surfing here is aware of preserving the web pages' annotation popularity, i.e. it prevents the case that the initially annotated pages have no annotation at all after propagation.
2) **Multiple-Annotation Surfing**: Most previous propagation tasks focus on only one value for each page. For example, HITS and PageRank propagate the quality value for each page. Models in [30] and [25] propagate the relevance value of a query for each page. Differently, propagation of social annotations needs to deal with a higher dimension, i.e. multiple annotations may be propagated to the same page. Moreover, during the propagation, different annotations should preserve their overall popularities on the whole corpus. Multiple-annotation surfing is proposed for above purposes.
3) **Multiple-Link-Type Surfing**: The annotation can also be propagated via different types of links. Multiple-link-type surfing models the general case of propagating annotations via different types of links with different a priori probabilities. In our implementation, the most widely used sitemap-tree links and hyperlinks are investigated.
4) **Constraint-Guided Surfing**: Different links have different capabilities for propagation. Constraint-guided surfing is proposed to guide the annotation propagation and avoid/ alleviate annotation's topic drifting. Specifically, we study the link connection strength. The links connecting two pages more closely should have a higher priority to be propagated with.

To evaluate the social propagation model, we build a dataset by randomly crawling 100 most popular English sites in Alexa as well as ten famous sites which are familiar to academic. The dataset consists of 40,422 pages distributed in 87 domains. 48,481 sitemap-tree links and 786,440 hyperlinks are extracted. Then each page is queried in Del.icio.us history[5] to check whether it is annotated. 20.31% of them are annotated with 44,515 distinct annotations.

We evaluate the proposed model with various propagation settings. Sitemap-tree links, hyperlinks and constraint are incrementally added based on basic surfing and multiple-annotation surfing. Experimental results show that the basic surfing is very effective in protecting the original annotations while improving the annotations' coverage. Multiple-annotation surfing works well in preserving different annotations' initial popularity. Multiple-link-type surfing can boost the social annotation effectively by sitemap-tree links

[5] http://del.icio.us/url/.

and hyperlinks. The constraint-guided surfing can further improve the propagation accuracy by alleviating annotation's topic drift.

The generated annotations have many applications, e.g., page classification, clustering and summarization. As case studies, we evaluate the proposed model by applying the propagated annotations to search and classification. The search experiments are carried out with 566 auto-generated queries. The experiments of classification are based on manual labeling of 16,493 pages into 12 top ODP categories. Experimental results on these applications show that propagated annotations bring significant improvement.

The rest of the paper is organized as follows. Section 2 discusses some related work. Section 3 formalizes the propagation problem. Section 4 describes social propagation model in detail. Section 5 presents the experimental results and Section 6 gives some discussions. Finally, we give some concluding remarks and future work in Section 7.

2 Related work

2.1 Research on social annotations

Existing research on social annotations includes "folksonomy" [12, 19, 21], semantic web [22, 34], search and browsing [1, 7, 18, 35, 40], personalization [5, 23], event detection, visualization [8, 27], etc.

Early research on social annotations focuses on "folksonomy" [12, 19, 21, 32]. "Folksonomy", a combination of "folk" and "taxonomy", was first proposed by T. V. Wal in a mailing list [31]. It provided user-metadata rather than the professional created or author created metadata [19]. In [21], P. Merholz argued that a folksonomy could be quite useful in that it revealed the digital equivalent of "desire lines". In [4], the authors analyzed the effectiveness of tags for classifying blog entries. [11] analyzed the structure of collaborative tagging systems as well as their dynamical aspects. Halpin et al. produced a generative model of collaborative tagging in order to understand the basic dynamics behind tagging [12]. Hotho et al. proposed Adapted PageRank and FolkRank to find communities within the folksonomy [14]. A general introduction of folksonomy could be found in [26] by E. Quintarelli.

Some applications based on social annotations have also been explored, e.g. semantic web [22, 34, 38], search and browsing [1, 7, 18, 35, 40], personalization [5, 23], event detection, visualization [8, 27], and Web page classification [15]. M. Dubinko et al. considered the problem of visualizing the evolution of tags [8]. They presented a new approach based on a characterization of the most interesting tags associated with a sliding time interval. P. Mika proposed a tripartite model of actors, concepts and instances for semantic emergence [22]. Wu et al. explored semantics from social annotations in a statistical way [34]. Zhou et al. further developed an unsupervised model for exploring the hierarchical semantics from the social annotations [38]. Li et al. improved the web page browsing experiences by using social annotations [18]. Dmitriey et al. used annotations to improve the quality of intranet search [7]. Bao et al. investigated the capability of social annotations in improving the quality of web search [1]. Xu et al. smoothed the language model for web search using social annotations [35]. Zhou et al. proposed a unified framework to combine the modeling of social annotations with the language modeling-based methods for information retrieval [40]. Noll and Meinel adjusted the web search for different web users via social bookmarking and tagging [23].

Dubinko et al. considered the problem of visualizing the evolution of tags [8]. Recently, Rattenbury et al. proposed an approach for extracting place and event semantics from tags [27]. The utilization of social annotations for Web page classification has also been addressed recently. For example, Kamishima, Hamasaki and Akaho developed a new algorithm BaggTaming, where social annotations were considered as wild data, and labels as tame ones, to improve the accuracy of Web page classification [15]. All above work was conducted with the assumption of having enough social annotations. However, the proportion of web resources with social annotations still keeps small. As a complementary, this paper proposes a new model to propagate the social annotations to the un-annotated web resources.

Some previous research has also been conducted on generating/extracting keywords for web pages [10]. Different from their work, we are aiming to enrich the annotations which consist of human knowledge. Another related work is Chirita et al.'s P-TAG, which produced keywords relevant to both its textual content and the data residing on the surfer's desktop [5]. In contrast with personalized tag, our approach is to propagate common understanding for web pages.

2.2 Propagation

Propagation is an idea on transferring some known-items to unknown-items. It is a simple, yet effective way in lightening the data sparseness problems. Many propagation approaches have been proposed, e.g., Spreading Activation [29], Manifold Propagation [39] and Label Propagation [41].

Propagation has been used in many applications, e.g., static quality propagation [2, 13, 16, 17, 24, 36], anchor text propagation [3], and relevance propagation [6, 9, 25, 30, 33]. Kleinberg proposed HITS algorithm for discovering hubs and authorities [16]. Page et al. proposed PageRank for popularity propagation [24]. In [36], Xue et al. further discovered site hierarchies for page quality estimation. Borges and Levene compared two quality ranking methods of web pages in a site, i.e. Site Rank, which was an adaptation of PageRank to the granularity of a web site, and Popularity Rank, which is based on the frequencies of user clicks on the outlinks in a page. Their experimental results showed that Site Rank provided a reasonable first order approximation of the aggregate behavior of users within a web site given by the Popularity Rank. Kumar et al. introduced a time graph for blog space and extended Kleinberg's notion of temporal bursts to derive bursty communities of blogs that were topically and temporally focused [17]. All their work is based on the idea of propagating the static quality from one page to another via web link and/or site hierarchy. Zhu and Ghahramani proposed label propagation for learning from labeled and unlabeled data [41]. Crestani and Lee proposed an association web search system WebSCSA for relevance propagation [6]. Flesca et al. propagated the user interest via the content and usage similarities among Web pages for personalized Website navigation [7]. Qin et al. proposed a generic relevance propagation framework, and then provided comparison study on the effectiveness and efficiency of various representative propagation models [25]. Recently, Shakery and Zhai proposed a general relevance propagation framework of combining content and link information [30].

The idea of propagation can also be used in lightening the sparseness of social annotations. However, as discussed before, the propagation of social annotations is different due to their particular properties. Consequently, we developed a new method to propagate social annotations effectively.

3 Problem formulation

To the best of our knowledge, no previous work has addressed the problem of social annotation propagation. To make it clear and easy for understanding, we formally define the social annotation propagation problem as follows:

As we can see from Figure 1, two types of objects (*web pages* and *annotations*) and two types of relations (*links* and *distribution*) are involved in social propagation. In some sense, social propagation is to change the *distribution* between *annotations* and *web pages* via *links* among *web pages*.

Before going on to the next section, we would like to make several remarks as below:

- While coverage is easy to measure, the effectiveness of property preservation is much harder to measure. Many properties of social annotations have been discussed before. In this paper, we focus on preserving three most representative properties of social annotations: 1) keyword property [1, 7, 18, 35], which means annotations are good summary of the corresponding web page; 2) popularity property, which indicates the amount of annotations represents a page's popularity [1, 8, 27]; and 3) complexity property [12, 18], which expresses the power-law distribution of social annotations.
- Increasing the coverage and preserving the properties of social annotations are contrary to each other. On one hand, a naive approach to increase the coverage is to propagate the annotations thoroughly via all available links. However, such propagated annotations would be useless or even harmful for social annotation based applications. On the other hand, a negative approach to preserve the novel properties is to do nothing. Obviously, it is useless in helping alleviate the data sparseness problem. Good leverage of both sides is the key to the effective propagation of social annotations, which is emphasized throughout this paper.

4 Social annotation propagation

In this section, we start with introducing our basic model which is able to protect page's original annotations to some extent. Then, we elaborate its extensions from different

Figure 1 Definition of social annotation propagation.

Given: A set of ***web pages*** $P = \{p_1, p_2, \ldots, p_m\}$, which are connected with a collection of ***links*** $L = \{l_1, l_2, \ldots, l_r\}$; for each $l \in L$, $l \in P \times P$;

A set of ***annotations*** $A = \{a_1, a_2, \ldots, a_n\}$ which are assigned to web pages P with a ***distribution*** of $D = \{A(p_1), A(p_2), \ldots, A(p_m)\}$; for each $A(p_i) \in D$, $A(p_i) = \{a_1(p_i), a_2(p_i), \ldots, a_n(p_i)\}$ where $a_j(p_i)$ means the count of annotation a_j that page p_i has.

Return: A new propagated distribution D' = ***SocialPropagation*** (P, L, A, D), such that:

1. Increases the coverage of web pages with annotations.
2. Preserves the original properties of social annotations.

perspectives, including multiple annotations, multiple link types and propagation constraints. All the proposed models are based on the Random Surfer.

4.1 Social propagation—basic model

Let's first consider a simplified configuration of propagating single social annotation via the same type of links. Assume that there is only one annotation a assigned to a page p. The purpose of a web surfer is to collect more pages related to annotation a. Then it may have a Markovian random surfing process starting at p as follows: (Figure 2)

In the above model, a is similar to the expansion factor of PageRank [3, 24]. It indicates the portion of annotation for propagation at each step. Then the probability of a web page p owning annotation a at step $n+1$ can be derived as:

$$a^{(n+1)}(p) = (1-\alpha)a^{(n)}(p) + \alpha \sum_{p' \to p} a^{(n)}(p') \frac{|L(p' \to p)|}{|L(p' \to *)|}, \tag{1}$$

where $a^{(n)}(p)$ means the probability of annotation assigned to page p at step n; $L(p' \to p)$ means the collection of links pointing from page p' to page p and $L(p' \to *)$ means the collection of links pointing from p'. Assuming the random surfer runs for enough long period, the above iteration will reach an equilibrium distribution [13].

The above model is able to propagate social annotations to un-annotated pages, i.e. improving the coverage of annotations. However, another problem arises that some pages initially annotated may lose all their annotations after the propagation. Such a random surfing can not guarantee the initial popularity of web pages.

Figure 3 illustrates a simple case. Assume that there are three pages p_0, p_1 and p_2, and two links $p_0 \to p_1$ and $p_1 \to p_2$. Before the propagation, only page p_0 is annotated with annotation a. Then, we apply the Eq. 1 iteratively with $\alpha>0$. Finally, we find that the annotation of p_0 is propagated to page p_2 via page p_1 completely. We call this phenomenon as over propagation.

To avoid over propagation, we introduce a new step into the naive model, i.e. the random surfer has an opportunity to stop with probability β and to continue with probability (1-β). Then the naive surfing can be refined as follows: (Figure 4)

In the basic model, β is a damping factor which preserves the web pages' original annotations to some extent by affecting the iteration steps. The final expected distribution becomes:

$$a(p) = \sum_{i=1}^{\infty} \beta^i a^{(i)}(p). \tag{2}$$

In the following sections, for simplicity, we focus on extending the naive model from different aspects to fit the requirement of effective social annotation propagation.

Figure 2 Random walk of naive model.

1. Stay at the same page with probability (1-α);
2. Move to a new page by following a link from one page to another with probability α.

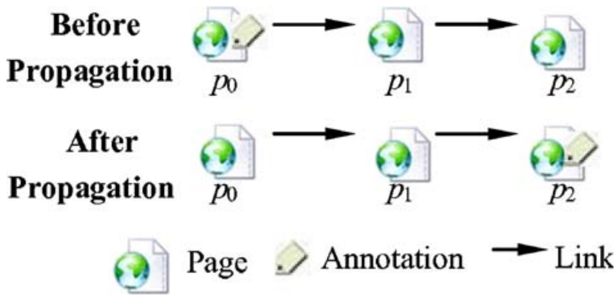


Figure 3 Illustration of over propagation.

4.2 Social propagation—multiple annotations

Now, let's consider a more complicated case of social annotation propagation. Assume that there is a set of annotations $A=\{a_1, a_2,\ldots, a_n\}$. Before propagation, we have a count of c_i for each annotation a_i. In total, there are c annotations. Correspondingly, the random surfer has to decide which annotation to propagate. Figure 5 illustrates the new random surfing process. At first, each annotation a_i has probability c_i/c to be randomly selected. Then, the surfer follows the naive model as described in Figure 2.

Note that above random walk preserves each annotation's popularity, i.e. the total count c_i is not changed for annotation a_i during the whole propagation process. Then we have the multiple-annotation propagation model as:

$$\overrightarrow{A^{(n+1)}}(p) = (1-\alpha)\overrightarrow{A^{(n)}}(p) + \alpha \sum_{p' \to p} \overrightarrow{A^{(n)}}(p') \frac{|L(p' \to p)|}{|L(p' \to *)|}, \tag{3}$$

where $\overrightarrow{A}(p)=\{ a_1(p), a_2(p),\ldots, a_n(p)\}$ is a vector that stores the probabilities of different annotations assigned to page p.

4.3 Social propagation—multiple link types

Similar to previous work on relevance propagation [30], the annotations can also be propagated via different types of links, e.g., hyperlinks and sitemap-tree links. Our model can be further extended to propagating social annotations through different kinds of links. Without loss of generality, we assume that there are l types of links $L_T=\{L_1,L_2,\ldots,L_l\}$. For

Figure 4 Random walk of basic model.

0. Terminate surfing with probability β or continue with probability (1-β);

If continue:

1. Stay at the same page with a probability (1-α);
2. Move to a new page by following a link from one page to another with a probability α.

Figure 5 Random walk with multiple annotations.

1'. Random selection of one annotation for propagation;

1. Stay at the same page with a probability (1-α);

2. Move to a new page by following a link from one page to another with a probability α.

each link, the surfer has different probability $P(L_i)$ to follow according to the link's type. Then, the naive random surfer's walk can be adjusted as follows: (Figure 6)

The probabilities of a page annotated by surfer are transformed as follows:

$$a^{(n+1)}(p) = (1-\alpha)a^{(n)}(p) + \alpha \sum_{L_i \in L_T} P(L_i) \sum_{p' \to p \mathrm{via} L_i} a^{(n)}(p') \frac{|L_i(p' \to p)|}{|L_i(p' \to *)|}, \qquad \sum_{i=1}^{l} P(L_i) = 1, \tag{4}$$

where " $p' \to p$ via L_i " means page p' and page p are linked with link type L_i.

4.4 Social propagation—constraint

Links are good indicators for propagation, but they are not the whole story. The connection strength of two linked pages varies a lot. For example, some links are given to pages describing other topics. To prevent/alleviate the topic drifting, we add a constraint factor $P(p' \to p)$ to guide the surfer's walk. Figure 7 shows the details.

The corresponding revised model with propagation constraint is shown in Eq. 5.

$$a^{(n+1)}(p) = (1-\alpha)a^{(n)}(p) + \alpha \sum_{p' \to p} a^{(n)}(p') \frac{|L(p' \to p)|}{|L(p' \to *)|} P(p' \to p), \qquad \sum_{p|p' \to p} \frac{|L(p' \to p)|}{|L(p' \to *)|} P(p' \to p) = 1. \tag{5}$$

In this paper, we propose to estimate the link connection strength for propagation as Eq. 6.

$$P_i(p' \to p) \propto \frac{\overrightarrow{p'} \cdot \overrightarrow{p}}{\left|\overrightarrow{p'}\right| \left|\overrightarrow{p}\right|}, \tag{6}$$

where $\overrightarrow{p'}$ and $\vec{p}$ are TFIDF vectors of pure page contents.

4.5 General model

The random surfing based models are quite flexible. The above surfing steps can be merged and we present our general social surfing process as follows: (Figure 8)

Figure 6 Random walk with multiple link types.

1. Stay at the same page with a probability (1-α);

2'. Select a type of links L_i with probability $P(L_i)$;

2. Move to a new page by following a link of L_i with a probability α.

Figure 7 Random walk with constraint.

1. Stay at the same page with probability (1-α);
2. Move to a new page by following a link from one page to another with a probability α; the selection probability of specific link $p' \rightarrow p$ is in direct proportion to $P\ (p' \rightarrow p)$.

Then, the final distribution of annotations can be calculated by Eq. 2, where the corresponding iteration formula is replaced as follows:

$$\overrightarrow{A^{(n+1)}}(p) = (1-\alpha)\overrightarrow{A^{(n)}}(p) + \alpha \sum_{L_i \in L} P(L_i) \sum_{p' \rightarrow p \mathrm{via} L_i} \overrightarrow{A^{(n)}}(p') \frac{|L_i(p' \rightarrow p)|}{|L_i(p' \rightarrow *)|} P(p' \rightarrow p),$$
$$\sum_{i=1}^{l} P(L_i) = 1, \tag{7}$$
$$\sum_{p|p' \rightarrow p \mathrm{via} L_i} \frac{|L_i(p' \rightarrow p)|}{|L_i(p' \rightarrow *)|} P(p' \rightarrow p) = 1.$$

5 Experimental results

5.1 Data setup

While the proposed model is generally applicable to the whole World Wide Web, we compile a pilot data set, which contains both Web site information and social tagging history for evaluation.

5.1.1 Web data setup

To evaluate the proposed propagation model, we manually build a data set by crawling the top 100 most popular English sites in Alex as well as ten famous sites which are familiar to academic. Some site examples are shown in Table 1.

As most of the popular sites have a huge number of pages, we crawled a portion of them in a breadth-first way. The detailed information of downloaded pages is shown in Table 2. Note that the number of domain is not equal to 110 because some of the top 100 English sites forbid crawling in their robots.txt.

5.1.2 Social data setup

For each collected page, we query Del.icio.us history[6] to check whether it is collected by the web users. In [4], the authors categorized the social annotations into three basic strategies for tagging: 1) Annotating information for personal use; 2) Placing information into broadly defined categories; and 3) Annotating particular articles so as to describe their content.

Given a page, the list of most popular tags is usually mixed with different categories of annotations described above. In this paper, we are mainly interested in propagating the 2nd and 3rd categories of social annotations and manually identify a list of personalized annotations e.g. *toread*, *todo*, to filter out the 1st category. By no means, the list would

[6] http://del.icio.us/url.

Figure 8 Random walk of general model.

0.Terminate surfing with probability β and continue with probability (1-β);

If continue

1'. Random selection of one annotation for propagation;

1. Stay at the same page with a probability (1-α);

2'. Select a type of links L_i with probability $P(L_i)$;

2. Move to a new page by following a link from one page to another with a probability α; the selection probability of specific link $p' \rightarrow p$ is in direct proportion to $P(p' \rightarrow p)$.

cover all personal annotations, especially with the fast developing of web and social annotation services. Although it is helpful to derive the rest personal annotations automatically, identification of personalized annotations is not trivial. In fact, even the manually collected personal tags may belong to the 2nd or 3rd category. For example, *Todo*, may be used to describe a product Todolist,[7] which is a popular online task manager.

Besides the personalized annotations, we further identify other useless annotations via several simple yet effective rules as follows: 1) System generated annotations, e.g. *system:undefined*; 2) Noisy annotations, e.g., the annotations appear less than five times in the whole corpus. The annotation statistics before/after filtering is given at Table 3.

5.1.3 Link data setup

Previous studies show that many kinds of links can be utilized in static propagation, relevance propagation etc. The top two popularly used links are sitemap-tree links and hyperlinks. Although our model is general enough to handle various kinds of links, without loss of generality, in this paper, we still focus on studying the propagation of annotations via these two kinds of links. The extracted sitemap-tree link and hyperlink information are shown in Table 4.

5.2 Experimental settings

For simplification, we represent the following settings as S1–S6, respectively. S1 only uses the text content of web pages. S2 makes use of both original page content and filtered manual annotations (Hereafter, manual annotations refer to original annotations assigned by web annotators). S3–S6 make use of both original page content and propagated annotations that derived from manual annotations with different propagation settings. Table 5 shows the detailed information.

5.3 Evaluation of propagation

To evaluate the effectiveness of social propagation, we apply the social propagation model with settings of S3–S6 separately. During the propagation, α which depicts the transmitting

[7] http://todolist.sourceforge.net/.

Table 1 Website examples.

Popular english sites	Famous academic sites
http://www.yahoo.com	http://trec.nist.org
http://www.youtube.com	http://www.berkeley.edu
http://www.live.com	http://www.harvard.edu
http://www.orkut.com	http://www.stanford.edu
http://www.facebook.com	http://www.mit.edu

probability of an annotation from one page to another, is set similarly to PageRank's expansion factor 0.85 [3]. β, which describes the terminating probability of propagation, is empirically set to 1-α, i.e. 0.15. In some sense, α and β are complementary with each other in controlling to what extent an annotation should be propagated. The default a priori probabilities of choosing sitemap-tree links and hyperlinks are empirically set to 0.4 and 0.6 respectively, based on their connection reliabilities, The models converge with a total difference of less than 0.001 after 80–110 iterations (One iteration here means a round of propagation with all the available links for each annotation in the collection).

5.3.1 Propagation coverage

Table 6 shows that both sitemap-tree links (S3) and hyperlinks (S4) improve the coverage of social annotations significantly, and hyperlink improves more. By combining the two types of links, S5 can further improve the coverage to 66.90%. The use of constraint does not affect the coverage a lot and get coverage of 66.53% (S6), which is about four times of the coverage of annotations before propagation.

Figure 9 shows the detailed annotation coverage improvement for web pages at different URL depths (i.e. the position of URL at the corresponding sitemap-tree). It is easy to see that S3–S6 all improve the coverage on each depth largely. We can also find that the improvement of sitemap-tree (S3) decreases with the increasing of depth. This is mainly because 1) the annotations of pages can not be directly propagated from top to bottom and 2) the annotations that are able to propagate decrease with the influence of dumping factor β. In contrast, with the help of hyperlinks (S4), the coverage of propagated annotations remains high even the manual annotations of deeper URL are much fewer.

5.3.2 Property preservation

Note again that a good social propagation should not only simply increase the coverage of social annotations, but also preserve the original properties of social annotations as well as possible. Here, we will evaluate the keyword property, popularity property and complexity property of propagated social annotations.

Table 2 Web page statistics.

Domain count	Host count	Path count	Page count
87	111	20,034	40,422

Table 3 Annotation statistics.

		Original	Filtered
Annotation	Total count	2,061,558	1,887,485
	Distinct count	44,515	2,812
Annotated page	Annotated count	8,209	5,742
	Annotated ratio	20.31%	14.21%

Keyword property Keyword property denotes whether the social annotations are good summary of web pages from the web users' perspective. Some case studies of web pages with different number of original annotations are given in Table 7. Note that result of social propagation here is not normalized for the ease of understanding. As we can see, the propagated annotations describe the target web page precisely and preserve the keyword property well. For the pages with large number of annotations, e.g. URL1 and URL2, propagation usually decreases their annotations. For pages with medium number of annotations, e.g., URL3, propagation may change their annotations amount according to their contexts. The pages with few annotations, e.g., URL4 and URL5, benefit from the social propagation most.

To evaluate the keyword property of the propagated social annotations quantitatively, we randomly selected 100 web pages which have more than ten distinct manual annotations and omitted their manual annotations during the propagation. Figure 10 shows the average overlap of top N annotations between original manual annotations and automatically propagated annotations. It is easy to see that the overlap of propagated and manual annotations is above 0.4 in most cases and S6 achieves the best performance at each level with the help of propagation constraint. Note that the propagated annotations not matched by manual annotations may also be useful in describing the target page from different perspectives as we have shown at Table 7, e.g., URL4 and URL5.

One may argue sometimes the social annotations can be over- propagated. For example, every Google webpage would have "Google" tag by propagation, which will be counted as the "right" annotation. However, it might not be helpful for ordinal users of social tagging who wants to find related Web pages they don't know. This problem can be addressed from two points of view. Firstly, our model introduces a terminating parameter β which alleviate the problem of over-propagation. So, not every Google webpage can be propagated with "Google". In fact, we do find many Google web pages whose most popular tag is not "Google" but "search", "web2.0", "tools" instead. Secondly, we introduce a content based constraint to guide the propagation. So if a Google web page is not discussing itself (e.g.,

Table 4 Link statistics.

	Link type	Count
Sitemap-tree links	between path and file	40,422
	between path and sub-path	8,059
Hyperlinks	without duplication	633,294
	with duplication	786,440

Table 5 Experimental setting information.

Setting		S1	S2	S3	S4	S5	S6
Web page content		√	√	√	√	√	√
Original annotation			√				
Propagated annotation	Sitemap-tree			√		√	√
	Hyperlink				√	√	√
	Constraint						√

discussing its competitors), its probability to be propagated with "annotation" can be further reduced.

Popularity property Popularity here means the number of annotations a web page has. It is quite useful in web search [1], hot event detection and visualization [8, 27]. To evaluate the effectiveness of preserving a page's popularity on a specific annotation, we randomly select 500 pages annotated with *google*. As shown in Figure 11, the annotation distribution of *google* changes after propagation. Generally, the annotation count decreases on pages with many annotations, while it increases smoothly on pages with few annotations. The popularity order of different pages is preserved in most cases, except some pages whose counts change greatly by propagating a great number of annotations from neighbor pages.

Complexity property Complexity of annotations has been discussed in many applications before [12, 18]. The complexity of social annotations results in power-law distribution of social annotations. Figure 12 shows the distribution of original annotations and propagated annotations with different settings. As we can see, annotation propagation based on random surfing can preserve the power-law distribution well. In fact the propagation in some sense is a smoothed redistribution of social annotations by "*borrowing*" some annotations from plentiful pages for poor pages.

5.4 Applications of propagated annotations

The foregoing experimental results have shown that propagation can significantly improve annotation coverage while preserving the original properties of social annotations at the same time. In this section, we will further verify the effectiveness of the proposed model by evaluating the propagated annotations in other applications. As we can see later, by using the propagated annotations, classification and search of web pages can be improved significantly over using only content information.

Table 6 Annotation statistics after propagation.

		S3	S4	S5	S6
Annotated page	Count	13,713	24,032	27,041	26,891
	Ratio	33.92%	59.45%	66.90%	66.53%

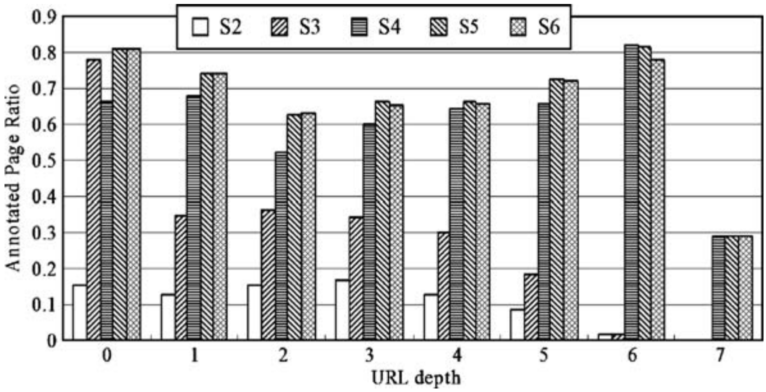


Figure 9 Annotation distribution(S2–S6) over pages with different URL depth.

5.4.1 Search

Previous studies show that social annotations are helpful in enterprise search [7], web search [1], personalized search [23] etc. Here we are to show the effectiveness of *propagated* annotations in web search.

System setup Our system uses the OKAPI BM25 [28] model to compute the similarity score between query and document content, and annotation field respectively. The term frequency component is implemented as follows:

$$TF(t,d) = \frac{(k+1) * f(t,d)}{k * ((1-b) + b * doclen/avgdoclen) + f(t,d)},$$

where $f(t,d)$ means the term count of t in document d. In the experiment, k and b are set to 1.2 and 0.75, respectively.

Query generation We automatically extract 566 queries and their corresponding ground truths from the ODP[8] as [1]. First, we merge the crawled data set with ODP, only URLs appear in both crawled data set and ODP will be preserved. Second, we extract the category paths as the query set and extract the corresponding web pages as the ground truths. Note that the term TOP in the category path is discarded. For example, the category path “TOP/Computers/Software/Graphics” would be extracted as the query “*Computers Software Graphics*”. Finally, we got 566 queries with 627 relevant documents. The average length of automatic queries is 7.647.

Evaluation metrics We evaluate the ranking algorithms over two popular retrieval metrics, namely *Mean Average Precision* (*MAP*), and *R-Precision*. Each metric focuses on one aspect of the search performance, as described below. MAP is defined as the mean of average precision over queries. R-precision is defined as the mean of top R_i precision. Here, R_i is the number of documents for the *ith* query in the ground truth.

[8] http://www.dmoz.org

Table 7 Case studies of propagated annotations.

URL1: http://code.google.com/index.html	
S2	google 3949, programming 2078, code 1876, api 1657
S6	google 958, programming 411, code 372, api 328
URL2: http://money.cnn.com/index.htm	
S2	finance 540, news 485, money 415, business 399
S6	finance 234, news 214, money 176, business 171
URL3: http://www.flickr.com/groups/moo/index.html	
S2	flickr 11, photography 5, design 3, tools 2
S6	flickr 11.4, photography 5.2, design 2.9, photos 2.3
URL4: http://www.adobe.com/products/photoshop/compare/	
S2	photoshop 2, adobe 1
S6	adobe 1.9, photoshop 1.8, photography 1.4, software 1.4
URL5: http://www.bbc.co.uk/schoolradio/music/index.html	
S2	not available
S6	bbc 0.33, radio 0.13, education 0.12, news 0.12

Search results Table 8 shows the search results using different resources, i.e. original content (S1), content with manual annotations (S2) and content with propagated annotations (S6). Both MAP and R-Precision are improved incrementally. The propagated annotations may further benefit the web search and achieve the best result which outperforms S1 by a relative improvement of 12.17% on MAP.

5.4.2 Classification

Classification may also benefit from the propagated annotations. To understand the effect of *propagated* propagation better, we are to show whether a page originally has no annotation can benefit from our propagation. So we make a direct comparison of S1 and S6 on classification. Here, S1 refers to a set of pages S_P which has no manual annotations at all. S6 refers to the same set of pages $S_p^{'}$ which contains propagated annotations only. More

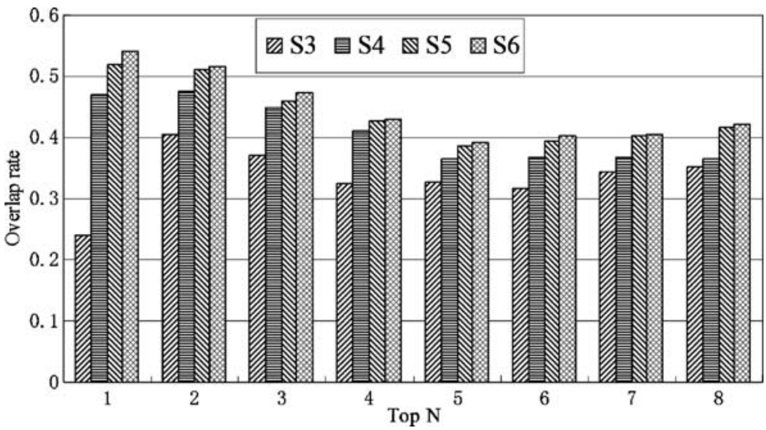
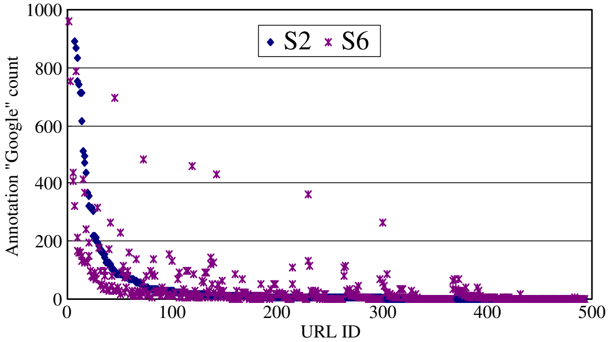


Figure 10 Average top N overlap of manual annotations and propagated annotations.

Figure 11 Pages' popularity on annotation "google".



specifically, we filter out 5,742 pages with manual annotation (as presented in Table 3). Then, a group of students are invited to manually assign a subset of web pages without manual annotations into categories defined by ODP. We finally get 16,493 pages with propagated annotations distributed in 12 ODP categories. Table 9 shows the distribution of each category. Perception [20] is used for classification.

Evaluation metrics To evaluate the effectiveness of classification results, we use the standard recall (r), precision (p) and F_1 measures. Recall is defined to be the ratio of correct assignments by the system divided by the total number of correct assignments. Precision is the ratio of correct assignments by the system divided by the total number of the systems' assignments. The F_1 measure defines as:

$$F_1(r,p) = \frac{2rp}{r+p} \tag{8}$$

The micro-averaging F_1 and macro-averaging F_1 are introduced [37] for measuring the average performance over all categories. The micro-averaged scores tend to be dominated by the performance of common categories, and the macro-averaged scores are influenced by the performance of rare categories.

Classification results We merge the text content and social annotations for classification. The best ratio to combine the social annotation and text content is 0.3 : 0.7, based on a training set of pages with manual annotations. We use the same ratio to combine the propagated annotations and page content. Table 10 shows the classification results on S1

Figure 12 Annotation distribution over all pages.

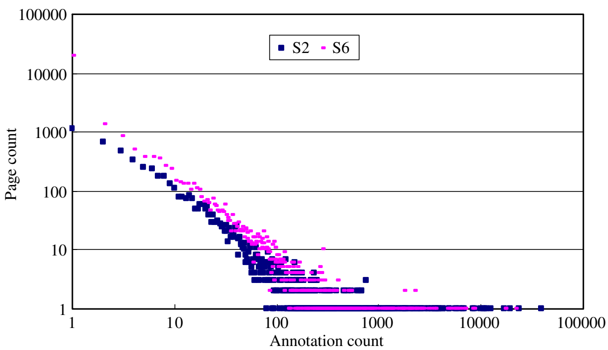


Table 8 Evaluation of search results with different settings.

	MAP	R-Precision
S1	0.3246	0.2344
S2	0.3577	0.2646
S6	0.3641	0.2730

and S6. S2 is not involved as a baseline here since the pages used for classification are all without manual annotations. All the results are evaluated based on 5-fold cross validation. It is easy to see that social propagation brings improvement on each measure. With the help of propagated annotations, S6 outperforms S1 by 18.37% and 5.87% on Micro-F1 and Macro-F1, respectively.

6 Discussion

6.1 Scalability of the propagation

Propagation of social annotations is different from previous propagation tasks, like static quality propagation and relevance propagation. Both relevance propagation and static quality propagation focused on a single value, not the annotation vectors, as social propagation does. From this point of view, propagation of annotations requires more computation. Assume that there are $|L|$ links and $|A_P|$ distinct annotations per page. For each iteration, the computation complexity is approximately $O\ (|L|\ |A_P|)$.

Two properties of social annotations can be further utilized to accelerate the propagation. 1) Distributed propagation: in our current model, the propagation of different annotations can be processed concurrently. Besides, it is also possible for partial propagation, which is especially useful in propagating some new hot annotations emerging on the web without changing any previous propagation results. 2) Incremental propagation: as addressed by [11, 12], the annotation will become stable for some popular pages. As a result, the incremented annotations of stable pages only change their popularity of previous propagated distribution and can skip step-by-step propagation.

Table 9 Category statistics.

Cate.	Num.
Computer	5,971
News	2,868
Business	1,941
Games	1,339
Reference	1,103
Arts	1,051
Recreation	1,028
Society	517
Sports	338
Shopping	149
Science	130
Health	58

Table 10 Classification results with different settings.

	Micro			Macro		
	Prec.	Recall	F1	Prec.	Recall	F1
S1	0.4730	0.5010	0.4844	0.7590	0.4245	0.5438
S6	0.5882	0.5602	0.5734	0.7796	0.4586	0.5757

6.2 Propagation through more links

Besides hyperlinks and sitemap-tree links, there are also many other link information available. For example, links can also be generated from manual web directories, like ODP. The web pages under the same categories usually share the same topic and as a result, the annotations can propagate via either links among siblings or links between parent and child pages. The proposed model is quite flexible and such links can be easily integrated. In this paper, we did not introduce the ODP based links yet since the connections represented by ODP are much looser and may bring more noises. However, the links derived from ODP can connect two sites/pages which may not be easily connected via hyperlinks or sitemap-tree links. We leave it as one of our future directions in further boosting social annotations.

Another direction is to propagate the social annotations via more fine-grained web links. Many methods can be used to refine the link information e.g., noise link removal and block level link detection. Noise link removal can be directly applied in our setting as a preprocessing step. As for block level links, the initial block-level annotation for propagation is required. The initial block-level annotation can be obtained either from page-level annotations via machine learning, or from block-level annotation services, e.g., CiteULike[9] and Technorati[10] which allow the assignment of annotations to objects within a page. We argue that the fine-grained web link based annotation propagation will produce better results.

6.3 Propagation with more constraints

As we have seen that the content based constraint does alleviate the topic drifting effectively. In fact, more constraints can be introduced in our propagation model. First of all, annotations have different capability for propagation. For example, some annotations only belong to the original web pages and are supposed not to be propagated. Secondly, propagation can be adjusted based on the mutual relationship among annotations [1, 22, 34]. Thirdly, propagation is also affected by the similarity between annotations and content of the target page. Furthermore, different annotations may suit for propagation via different links. Some annotations can be propagated well via sitemap-tree links while some others are more suitable for hyperlink based propagation. The above constraints can help avoid topic drifting better. However, they also introduce more computations and destroy the concurrency properties of current propagation model.

[9] http://www.citeulike.org/

[10] http://technorati.com/

6.4 Propagating more information

The social annotations are usually modeled as a quad-tuple, i.e.<*user*, *annotation*, *resource*, *time*>which means that a *user* gives an *annotation* to a specific *resource* at a specific *time*. In this paper, we focus on the propagation of annotations over resources since they are the most useful part of the social structures. However, the user and time information may also be useful in certain applications. For example, *user* information can be used for personalized search [23, 34], browsing [18] and tag generation [5]. The time information may be used for hot topic browsing [18], visualization [8] and detection [27].

As a general framework, it is easy to incorporate user and/or time information. The simplest way is to generalize the concept of annotation as constrained annotations, i.e. let constrained annotation be <annotation, user > or <annotation, user, time >. Then our random walk can be performed over the constrained annotations. Another way is to add a new step for the random surfer, e.g., the surfer may randomly select a user or a specific time period before selecting specific annotation for propagation. It would be interesting to see that user specific information can be extended via the propagation and support personalized search better. It would also be interesting to see whether more hot and detailed topics can be discovered from the propagated time-related annotations. We leave these as our further work.

7 Conclusion

Social annotations are novel resources as well as useful information in many applications, such as emergent semantics, search and browsing. Although they are developing fast, they only cover a small portion of the fast growing World Wide Web, and thus suffer from the sparseness problem. In this paper, we propose to boost the social annotations of web pages automatically using propagation. The main contributions can be summarized as follows:

1. The observation of social annotations' sparseness problem and proposal of automatically boosting social annotations using propagation.
2. The proposal of the general propagation model as well as the study of annotation and web link properties for effective propagation.
3. The extensive evaluation of the proposed model. Further evaluation shows that the propagated social annotations benefit various applications, like search and classification, significantly.

In the future, we are going to study more sophisticated features for social annotation propagation over web pages. Besides, we will also study the propagation of social annotations for other digital resources like blogs and images, by exploring some pseudo link information.

References

1. Bao, S., Wu, X., Fei, B., Xue, G., Su, Z., Yu, Y.: Optimizing web search using social annotation. In: Proc. of WWW, pp. 501–510 (2007)
2. Borges, J., Levene, M.: Ranking pages by topology and popularity within web sites. World Wide Web J **9**(3), 301–316 (2006). doi:10.1007/s11280-006-8558-y
3. Brin, S., Page, L.: The anatomy of a large-scale hypertextual web search engine. In: *Computer Networks and ISDN Systems*, **30**(1–7):107–117 (1998)

4. Brooks, C.H., Montanez, N.: Improved annotation of the blogosphere via autotagging and hierarchical clustering. In: Proc. of WWW, pp. 583–589 (2006)
5. Chirita, P.-A., Costache, S., Handschuh, S., Nejdl, W.: P-TAG: Large scale automatic generation of personalized annotation TAGs for the web. In: WWW, pp. 845–854 (2007)
6. Crestani, F., Lee, P.L.: Searching the web by constrained spreading activation. In: Information Processing and Management. **36**(4):585–605 (2000). July
7. Dmitriev, P.A., Eiron, N., Fontoura, M., Shekita, E.: Using annotations in enterprise search. In: Proc. of WWW, pp. 811–817 (2006)
8. Dubinko, M., Kumar, R., Magnani, J., Novak, J., Raghavan, P., Tomkins, A.: Visualizing tags over time. In: Proc. of WWW, pp. 193–202 (2006)
9. Flesca, S., Greco, S., Tagarelli, A., Zumpano, E.: Mining user preferences, page content and usage to personalize website navigation. World Wide Web J **8**(3), 317–345 (2005). doi:10.1007/s11280-005-1315-9
10. Gao, X., Murugesan, S., Lo, B.W.N.: A simple method to extract key terms. Int J Electron Bus 4(3/4). IJEB. doi:10.1504/IJEB.2006.010863 (2006)
11. Golder, S.A., Huberman, B.A.: Usage patterns of collaborative tagging systems. J Inf Sci **32**(2), 198–208 (2006). doi:10.1177/0165551506062337
12. Halpin, H., Robu, V., Shepherd, H.: The complex dynamics of collaborative tagging. In Proc. of WWW, pp. 211–220 (2007)
13. Henzinger, M.R., Heydon, A., Mitzenmacher, M., Najork, M.: Measuring index quality using random walks on the web. In: Proc. of WWW, pp. 213–225 (1999)
14. Hotho, A., Jaschke, R., Schmitz, C., Stumme, G.: Information retrieval in folksonomies: Search and ranking. In: Proc. of ESWC, pp.411–426 (2006)
15. Kamishima, T., Hamasaki, M., Akaho, S.: BaggTaming—Learning from wild and tame data. In: Proc. of ECML/PKDD2008 Workshop: Wiki, Blogs, Bookimarking Tools (2008)
16. Kleinberg, J.: Authoritative sources in a hyperlinked environment. In: Proc. of 9th Annual ACM-SIAM Symposium. Discrete Algorithms, pp. 668–677 (1998)
17. Kumar, R., Novak, J., Raghavan, P., Tomkins, A.: On the bursty evolution of blogspace. World Wide Web J **8**(2), 159–178 (2005). doi:10.1007/s11280-004-4872-4
18. Li, R., Bao, S., Fei, B., Su, Z., Yu, Y.: Towards effective browsing of large scale social annotations. In: Proc. of WWW, pp. 943–952 (2007)
19. Mathes, A.: Folksonomies—Cooperative classification and communication through shared metadata. http://www.adammathes.com/academic/computer-mediated-communication/folksonomies.html, December (2004)
20. Mitchell, T.: Machine learning. McGraw-Hill (1997)
21. Merholz, P.: Metadata for the masses. October 19. http://www.adaptivepath.com/publications/essays/archives/000361.php (2004)
22. Mika, P.: Ontologies are us: A unified model of social networks and semantics. In: Proc. of ISWC, pp. 522–536 (2005)
23. Noll, M., Meinel, C.: Web search personalization via social bookmarking and tagging. In: Proc. of ISWC, pp. 367–380 (2007)
24. Page, L., Brin, S., Motwani, R., Winograd, T.: The PageRank citation ranking: Bringing order to the web. Technical report, Stanford Digital Library (1998)
25. Qin, T., Liu, T.-Y., Zhang, X.-D., Chen, Z., Ma, W.-Y.: A study of relevance propagation for web search. In: Proc. of SIGIR, pp. 408–415 (2005)
26. Quintarelli, E.: Folksonomies: Power to the people. Paper presented at the ISKO Italy-UniMIB meeting. http://www.iskoi.org/doc/folksonomies.htm (2005). June
27. Rattenbury, T., Good, N., Naaman, M.: Towards automatic extraction of event and place semantics from flickr tags. In: Proc. of SIGIR, pp. 103–110 (2007)
28. Robertson, S.E., Walker, S., Hancock-Beaulieu, M., Gull, A., Lau, M.: Okapi at TREC. In: Proc. of TREC, pp. 21–30 (1992)
29. Rumelhart, D., Norman, D.: Representation in memory. Technical Report, Department of Psychology and Institute of Cognitive Science, UCSD La Jolla, USA (1983)
30. Shakery, A., Zhai, C.: A probabilistic relevance propagation model for hypertext retrieval. In: Proc. of CIKM pp. 550–558 (2005)
31. Smith, G.: Atomiq: Folksonomy: social classification. http://atomiq.org/archives/2004/08/folksonomy_social_classification.html (2004). Aug 3
32. Wal, T.V.: Explaining and showing broad and narrow folksonomies. http://www.personalinfocloud.com/2005/02/ explaining_and_.html: (2005). February 21
33. Wang, J., Li, M., Li, Z., Ma, W.-y.: Learning ranking function via relevance propagation. Technical Report, Microsoft Research Asia (2005). November

34. Wu, X., Zhang, L., Yu, Y.: Exploring social annotations for the semantic web. In: Proc. of WWW, pp. 417–426 (2006)
35. Xu, S., Bao, S., Cao, Y., Yu, Y.: Using social annotations to improve language model for information retrieval. In: Proc. of CIKM, pp. 1003–1006 (2007)
36. Xue, G.-R., Yang, Q., Yu, Y., Zeng, H., Chen, Z.: Exploiting the hierarchical structure for web link analysis. In: Proc. SIGIR, pp. 186–193 (2005)
37. Yang, Y., Lie, X.: A re-examination of text categorization methods. In: Proc. of SIGIR, pp. 42–49 (1999)
38. Zhou, M., Bao, S., Wu, X., Yu, Y.: An unsupervised model for exploring hierarchical semantics from social annotations. In: Proc. of ISWC, pp.680–693 (2007)
39. Zhou, D., Weston, J., Gretton, A., Bousquet, O., Schölkopf, B.: Ranking on data manifolds. In: Proc. of NIPS (2003)
40. Zhou, D., Bian, J., Zheng, S., Zha, H., Giles, C.L.: Exploring social annotations for information retrieval. In: Proc. of WWW, pp. 715–724 (2008). [1]
41. Zhu, X., Ghahramani, Z.: Learning from labeled and unlabeled data with label propagation. Technical Report 02-107, CMU-CALD (2002)