Artif Intell Rev (2011) 36:139–151
DOI 10.1007/s10462-011-9207-0

Topic-based ranking in Folksonomy via probabilistic model

Yan'an Jin · Ruixuan Li · Kunmei Wen · Xiwu Gu · Fei Xiao

Published online: 9 February 2011

Abstract Social tagging is an increasingly popular way to describe and classify documents on the web. However, the quality of the tags varies considerably since the tags are authored freely. How to rate the tags becomes an important issue. Most social tagging systems order tags just according to the input sequence with little information about the importance and relevance. This limits the applications of tags such as information search, tag recommendation, and so on. In this paper, we pay attention to finding the authority score of tags in the whole tag space conditional on topics and put forward a topic-sensitive tag ranking (TSTR) approach to rank tags automatically according to their topic relevance. We first extract topics from folksonomy using a probabilistic model, and then construct a transition probability graph. Finally, we perform random walk over the topic level on the graph to get topic rank scores of tags. Experimental results show that the proposed tag ranking method is both effective and efficient. We also apply tag ranking into tag recommendation, which demonstrates that

Y. Jin · R. Li (✉) · K. Wen · X. Gu
School of Computer Science and Technology, Huazhong University of Science and Technology,
Wuhan 430074, China
e-mail: rxli@hust.edu.cn

Y. Jin
e-mail: jin.yan.an@smail.hust.edu.cn

K. Wen
e-mail: kmwen@hust.edu.cn

X. Gu
e-mail: guxiwu@hust.edu.cn

Y. Jin
School of Information Management, Hubei University of Economics,
Wuhan 430074, China

F. Xiao
Department of Software Technology, Wuhan Vocational College of Software and Engineering,
Wuhan 430074, China
e-mail: xfyaya@whvcse.com

the proposed tag ranking approach really boosts the performances of social-tagging related applications.

Keywords Tag ranking · Probabilistic model · Random walk · Tag recommendation

1 Introduction

Social tagging has recently emerged as a popular way to allow users to annotate and categorize web resources by assigning one or more descriptive words, which are referred to "tags". Its immediate success is contributed to low technical barriers and ease of use. Users can freely use any tags (words) to tag their favorite resources.

The assigned tags can benefit many applications. For example, Adam Mathes argues that tags as user-generated metadata are usually to facilitate some organization and access of information (2004). Bao et al. (2007a) utilize tags to optimize web search and Tang et al. (2009) try to learn an ontology from the freely authored tags. However, most existing works ignore an important problem: *tagging quality*. Preliminary statistics on a Bibsonomy data set[1] show that 62.7% of the tags are used by only 1–3 user(s) and only 7.4% of the tags are used by more than 10 users. The quality of the tags varies largely depending on the authors expertise. Therefore, there is a clear need to measure the quality of social tags, which is referred to as social tag ranking in this paper. In other words, the quality of the tags in the front of ranking list is higher than those in the end.

The tag ranking problem is non-trivial. For some topics, a tag may be very important while on some others, it may be not. For example, the tag "*tiger*" can refer to a creature, but it can also refer to the famous golfer *Tiger Woods* and Mac OS X v10.4 *Tiger*. The importance of the tag with respect to three different meanings (topics) would be quite different. When a user searches for a document on http://Delicious.com[2] and his intention is to find some information about Mac OS, the tag "*tiger*" would be very useful to help identify the information. However, if he wants to get much about the animal, the tag "*tiger*" will be not so useful. How to differentiate the importance of a tag on different topics is thus a challenging problem.

However, little research has been done on social tag ranking conditional on topics. It seems that tag ranking is no need for current social tagging systems. There are no explicit systematic guidelines and no scope notes in these systems. Users may not purposely deliberate how to choose tag words and tend to pick words out from their completely uncontrolled vocabulary. But this is not always true. If we want to leverage tags, tag ranking is important. For example, when it comes to recommending tags, we must determine top k relevant candidates. PageRank Brin and Lawrence (1998) is a most famous algorithm for computing the importance score of a web page. We also can use it to calculate the tags' authority. But to rank conditional on topics, traditional incoming flows should be split across different topics instead of being mixed together indiscriminately, as a tag spans multiple topics. The effect is to separate the authority score into a vector to record a tag's reputation with respect to different topics.

In this paper, we focus on a probabilistic topic model to rank tags with topic relevancy. We first extract topics for every resource from tag space using Latent Dirichlet Allocation (LDA) Blei and Ng (2003), a probabilistic model. Then we perform random walk over the topic level on the co-occurrence of users and resources graph to get the topic scores. Finally,

[1] http://www.kde.cs.uni-kassel.de/ws/dc09/dataset.

[2] The world's leading social bookmarking service.

we apply tag ranking into tag recommendation. Experimental results show that our proposed approach can significantly improve the performance.

The contributions of the paper are:

- A novel method incorporating topical features into PageRank without affecting their global properties, while providing insight into the topic-level transition within the global authority propagation.
- An extensive experimental comparison of our approach to a number of well-known ranking algorithms to show the superiority of our approach.
- A new mechanism for tag recommendation. A topic-sensitive tag will be labeled for a given resource following topic-specific authority.

The remainder of this paper is organized as follows: the background and related work will be introduced in Sect. 2. The topic level tag rank model is then explained in Sect. 3, with focus on how to construct topics from tags and a random walk on co-occurrence graph. The experimental results will be shown in Sect. 4. We conclude with discussions and future work in the last section.

2 Prior work

There are currently many studies about social tag space a.k.a folksonomy. General overviews on folksonomy systems and their strengths and weakness are given in http://www.adammathes.com/academic/computer-mediated-communication/folksonomies.html, Tony et al. (2005). Recently, works on more specialized topics such as user's tagging motivation (Korner et al. 2010), tag recommendation (Hotho et al. 2006), document recommendation (Guan et al. 2010), improving search performance Bao et al. (2007b) have been presented. The literature concerning the fundamental problem of ranking is still sparse. But it has some unilateral related work about ranking, such as FolkRank of Tag (Hotho et al. 2008), some even consider the issue of topicality (Haveliwala 2002; Nie et al. 2006).

2.1 Ranking in folksonomy

PageRank has been applied on web pages for effective and efficient information retrieval problem solutions. However, PageRank cannot be applied directly on folksonomy. This is contributed to two reasons as following: (a) the documents consist of short text snippets only, ordinary ranking schemes such as TF/IDF are not feasible. (b) the undirected triadic hyperedges of folksonomy, compared to the directed binary edges of web graph. Hence, Hotho et al. (2008) proposed a ranking algorithm for folksonomy, named FolkRank. They considered the structure of folksonomy as a factor which influences rank score of tags and converted hypergraph in folksonomy space into an undirected tripartite graph, then an adapted PageRank algorithm was not directly used on the tripartite graph. But we consider that tripartite graph may include much noise for ranking one component such as tag. Further, topic level ranking was not taken into account in FolkRank. Liu et al. (2009) proposed a tag ranking scheme for automatically ranking tags of a given image on Flickr website. They firstly estimated initial relevance scores for tags based on probability density estimation, then constructed a tag similarity graph for random walk. Finally they got a score for every tag for a given image.

2.2 Topic-sensitive PageRank

To our best knowledge, topic level ranking has not been introduced in folksonomy, but some literatures talked about it on web page. Traditional random walk methods only use one single score to measure a page's importance without considering what topics are talked about in the page content. As a result, the pages with a highly popular topic may dominate pages of other topics. An ideal solution might be that the topics talked about in pages should be considered and the pages should be ranked according to their different topics. With such a topic-based ranking score, different topic-based ranking lists with different topics will be returned.

Both Haveliwala (2002) and Nie et al. (2006) gave their solutions at topic level. Haveliwala (2002) computed a set of PageRank vectors biased a set of representative topics and generated more accurate rankings than with a single, generic PageRank vector. In contrast, Nie et al. (2006) proposed a topical link analysis model that formalizes this intuition by using a topic distribution to embody the context in which a link is created and thus affect the authority propagation. They showed that the surfer may take any one out of the following three atomic action: jumping to a random page and focusing on a random topic in the target page (action JJ); following a hyperlink and staying in the same topic (action FS); following a hyperlink and jumping to any topic in that page (action FJ). Then the probability that the surfer is on a page for a topic can be calculated. Topics in Haveliwala and Nie were both manually abstracted from ODP (Open Directory Project.)[3] Yang et al. (2009) also conducted a research on the problem of topic-level random walk. They proposed a four-step approach for topic-level random walk, employed a probabilistic topic model to automatically extract topics from documents and then we perform the random walk at the topic level. But as mentioned above, it cannot be directly applied into folksonomy.

In this paper, we focus on topic-dependent random walk in folksonomy. We propose a tag ranking approach in which tags can be automatically ranked according to their relevance. First, we extract all topics in tag space through LDA Blei and Ng (2003). Then, the authority score following random walk can be calculated on the topics relationship graph. Finally, we take tag recommendation as our application and evaluate it.

3 Topic-based tag ranking

In this section, we will introduce our topic-based tag ranking approach. We firstly give the definition of topic-based ranking in folksonomy, and then overview our approach, and finally introduce the probabilistic relevance score estimation and random walk-based refinement in detail.

3.1 Problem formulation

The goal of this work is to learn tag ranking model from user tagged resources. First we introduce some terminology. A social tagging system consists of users $u \in U$, tags $t \in T$, and resources $r \in R$. We call an annotation of a set of tags to a resource by a user a post. A post is made up of one or more triples (r_i, t_j, u_k). Thus, for learning the tag ranking model, we have the training data set $D = \{(r_1, t_1, u_1), \cdots, (r_n, t_n, u_n)\}$, t_i is a set of tags annotated to resource r_i by user u_i. For simplicity, we only consider resource as tag sets in this paper,

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

thus each resource is represented as a set of tag $r_i = \{t_i\}$. Hence, we extract topics from a resource through tags of the specific resource.

In fact, a user does not deliberate the ranking of tags conditional on various topics for future management and retrieval when he makes a post. But, obviously, the order conditional on various topics is valuable. Our overall goal is to order tags t with a given topic vector z$(z_1, z_2, \cdots, z_k)$, which are induced from all resources. From the standpoint of probabilistic, we may get the probability of tags conditional on the specific topic, and then sort them in descending order.

3.2 Overview of approach

The basic idea of topical link analysis is to incorporate topic distribution into the representation of each tag as well as the importance score of tag. Therefore, there are two vectors associated with each tag: the topic vector and the authority vector.

In social networks, a tag usually has interests on multiple topics. Formally, each tag $t \in T$ is associated with a topic vector $\theta_t \in R^T$ of T-dimensional topic distribution. Each element $\theta_{t_i z_k}$ i.e. $p(t_i|z_k)$ is represents the relative contribution from topic z_k with tag t_i as a whole. As shown in Fig. 1, a tag t_i's topic is represented by the corresponding topic vector in the topic level. This vector is normalized such that the sum of the probabilities is 1 ($\sum_z \theta_{t_i z_k} = 1$). This vector is static and solely determined by the tag set of all resources. In contrast, we assign each tag $t \in T$ an authority vector φ_t to measure its importance, where $\varphi_{t_i z_k}$ denotes tag t_i's importance score on topic z_k. This vector is obtained from the proposed topical ranking algorithm, which is dynamic during authority propagation. From Fig. 1, we can tell that the summation $\varphi_t = \sum_z \theta_{t_i z_k}$ is identical to the original non-topical importance score, e.g., the score obtained by PageRank algorithm. Meanwhile, a tag's authority distribution may differ from its topic distribution in topic level, and the transition between tags under different topic will infect authority score.

The authority vector of tags can be used to many applications. In our work, we use it for tag recommendation. When the authority vector for each tag is ready, we can train the given resource r_i and get the topic distribution ($\mu_{r_i z_k}$) of it, then we can recommend tag by $\mu_{r_i z_k} * \varphi_{t_i z_k}$. So far, we presented an overview of topic-based tag ranking approach. In the following, we give three main steps in Sect. 3.3 in detail: 1) use fundamental topic model LDA to get the topics in all resources; 2) construct a tag relationship graph for running random

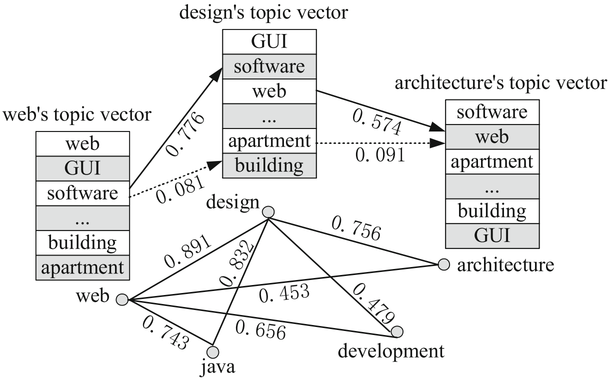

Fig. 1 Illustration of topics within tags

Table 1 Three example topics extracted from the dataset

Java development		Social tagging		Web design	
Tag	Prob.	Tag	Prob.	Tag	Prob.
Java	0.3158	Delicious	0.1806	Design	0.1978
Programming	0.0572	Socialtag	0.0797	Usability	0.0544
Eclipse	0.0505	Bookmarking	0.0675	Interface	0.0309
Develop	0.0415	Social	0.0675	Webdesign	0.0302
Apache	0.0224	Bookmarks	0.0499	Hci	0.0178
Framework	0.0193	Bookmark	0.0289	Patterns	0.0174
Development	0.0191	SNS	0.0249	Accessibility	0.0171
Gui	0.0178	Onlineservices	0.0218	Uml	0.0167
Frameworks	0.0131	Tagging	0.0186	Architecture	0.0099
Plugin	0.0105	Tags	0.0183	Pattern	0.0092

walk; 3) conduct random walk on the tag relationship graph and get the authority score of tags for various topic.

3.3 Topic-based tag rank model

3.3.1 Topics extraction

LDA is an unsupervised learning method widely used to model topics. LDA is a generative probabilistic model in which a document is generated by picking a distribution over topics, and given this distribution, picking the topic of each specific word. Then words are generated given their topics. In the context of tagging systems, the topics reflect a collaborative shared view of the resource, and the tags of the topics reflect a common vocabulary to describe the resource. Hence, we use $\boldsymbol{T}$, not document collection, as our input to run LDA model.

In our simulations, LDA models topics probability distribution over tags and resources being composed of multiple topics. Its parameters (the topic-tag and resource-tag distributions) can be estimated using an approximation technique known as Gibbs sampling. Gibbs sampling iterates multiple times over each tag t_i, and samples a new topic j for the tag based on the conditional probability $\mathrm{P}(z_k = j|t_i, z_{-k})$, where z_{-k} represents all topic-word and resource-topic assignments except the current assignment z_k for tag t_i. After the LDA model parameters converge, we can get the topic vector θ_t. In Table 1, we list three example topics which are extracted out of 100 topics by Gibbs sampling algorithm after 1,000 iterations. The extracted topics are *java development*, *social tagging* and *web design*.

3.3.2 Tag graph construction

In folksonomy, the structure is triadic context and differs from the web link structure. In order to gain superior tag ranking performance over FolkRank Hotho et al. (2008), we convert folksonomy into an undirected isomorphic graph and apply a random walk algorithm to the graph. In our approach, we define the tag graph as $\boldsymbol{G} = (\boldsymbol{V}, \boldsymbol{E}, \boldsymbol{W}, \Gamma)$, where

- $\boldsymbol{V} = \{t_1, \ldots, t_j\}$ is a set of nodes, which only consist of tags rather than users and resources.

- $\boldsymbol{E} = \{(t_i, t_j) | t_i \in \boldsymbol{V}, t_j \in \boldsymbol{V}\}$ is a set of edges. For simplicity, we use an weighted edge between nodes associated with the same resource while their co-occurrence number is greater than a threshold.
- $\boldsymbol{W}$ is the weight sets of $\boldsymbol{E}$. Its element w_{ij} is defined as: $w_{ij} = \frac{\boldsymbol{N}_{(t_i,t_j)}}{\sum_{kl}^{|T|} \boldsymbol{N}_{t_k,t_l}}$, where $\mathbf{N}(t_i, t_j)$ is the co-occurrence number of t_i and t_j.
- Γ is a $|\boldsymbol{T}|$ dimensional set. Its element is a $|\boldsymbol{Z}| * |\boldsymbol{Z}|$ transition matrix $\boldsymbol{M}$, whose element m_{ij} is the probability of tag t_i jumping to t_j following a specific topic. We define transition probability between tags as: $m_{ij} = p(t_j, z_k | t_i, z_l) = \frac{N(t_{ik} \rightarrow t_{jl})}{\omega_{ij}}$.

3.3.3 Topic random walk over tag graph

A topic random walk differs from the traditional PageRank algorithm due to its sensitivity on topics. A topic random surfer may label the tag t_i under the given topic z_k for a newly arriving resource. When choosing the next tag, he may either stay on the current tag with probability α or follow an outgoing link on any associated tag with probability $1 - \alpha$.

However, two scenarios must be considered when following an outgoing link. One scenario is topic stay, "TS", in which he is likely to stay on the same topic z_k to maintain topic continuity with a probability β. Another is topic-jump, "TJ", in which he may jump to any other topic z_l in the target tag with a probability $1-\beta$. As shown in Fig. 1, the tag *design* links to the tag *architecture* with a probability 0.574 under the same topic *web*, while it transits to the tag *architecture* with a probability 0.091 under a different topic. The probability of "TS", the solid line from the tag *design* to the tag *architecture* with the same topic *web*, is $(1 - \alpha)\beta\theta_{design,web \rightarrow web}$. When taking a "TJ" action, the preference among topics is determined by the topic in the target tag. The probability of "TJ", the dot line from *design* to *architecture* with different topics, is $(1 - \alpha)(1 - \beta)\theta_{design,apartment \rightarrow web}$.

In summary, at each step of random walk, the surfer may take either one out of the following two atomic actions: staying on the same topic tag (action TS), or jumping to any random topic tag (action TJ). Thus, the surfer's behavior can be modeled by probabilities:

$$\begin{cases} \text{TS action} : \mathrm{P(TS|t_i, z_l)} = (1 - \alpha)\beta \mathrm{w_{ij}} \\ \text{TJ action} : \mathrm{P(TJ|t_i, z_k)} = (1 - \alpha)(1 - \beta)\mathrm{w_{ij}} \end{cases} \tag{1}$$

However, how to select a topic, to stay on current topic or to jump to another topic? Essentially, the decision to whether to keep a topic is usually dependent on the topic of the current tag, i.e. $\theta_{t_j z_k}$, and the topic transition probability. If this tag is irrelevant to the topic of interest, the surfer is more likely to shift interest to another topic when entering a new tag. Hence, the probability to arrive at topic z_l in target tag t_j by the above actions can be described as:

$$\begin{cases} P(t_j, z_l | t_i, z_l, TS) = \theta_{t_i z_k} m_{kk} \\ P(t_j, z_l | t_i, z_l, TJ) = \theta_{t_i z_k} m_{lk} \end{cases} \tag{2}$$

The model can be used to compute the probability that the surfer is on tag t_j under the topic z_l as:

$$\begin{aligned} \varphi_{t_i z_l} = & \sum_{i:i \rightarrow j} P(t_j, z_l | t_i, z_l, TS) P(TS | t_i, z_l) \varphi_{t_i z_l} \\ & + \sum_{i:i \rightarrow j} \sum_{k \in Z} P(t_j, z_l | t_i, z_l, TJ) P(TJ | t_i, z_k) \varphi_{t_i z_k} \end{aligned}$$

$$= (1-\alpha)\beta \sum_{i:i\rightarrow j} \theta_{t_i z_k} w_{ij} m_{kk} \varphi_{t_i z_l} \tag{3}$$
$$+(1-\alpha)(1-\beta) \sum_{i:i\rightarrow j} \theta_{t_i z_k} w_{ij} \sum_{k\in Z} m_{lk} \varphi_{t_i z_k}$$

After the propagation converges, $\varphi_{t_j z_l}$ is the authority score of tag t_j on topic z_l.

4 Experiments

4.1 Dataset

Like many social tagging systems, Delicious.com allows user to manage a personal collection of links to the websites and describe those links with one or more words called tags. It is a web-based system that allows user to share bookmarks with each other. We crawled www.delicious.com and got a subset of it from Apr. 30 to Jun. 15 2009. There are 2,26,784 tags, 1,54,255 users, 3,53,016 resources and 11,16,996 posts in total. We chose a subset which also occurs in the top level category from ODP and then preprocessed dataset by (a) removing stop-words and punctuations, (b) lower-casing the obtained words, tags, user names and (c) removing words and tags that appear less than three times in the dataset. Statistically, there are $|T| = 47,687$ tags, $|U| = 18,273$ users, $|R| = 21,011$ resources.

4.2 Experimental settings

To compare our proposed approach, we choose well-known ranking algorithms as our baseline. These algorithms include traditional PageRank Brin and Lawrence (1998), FolkRank Hotho et al. (2008), topic-sensitive PageRank (TSPR) Haveliwala (2002) and topic link analysis for web search (TLA) Nie et al. (2006). We run them on the same link graph. The nodes of graph are preprocessed tags in Sect. 4.1, and nodes are joined with an edge when two tags occurred in the same resource greater than a larger times. But the formulas of computing authority score are completely different. For PageRank, we calculate the PageRank score and set damping factor d = 0.15 given by Brin and Lawrence (1998). For FolkRank, we use the equation described in Hotho et al. (2008) to get the rank of tags in folksonomy and set parameters like in Hotho et al. (2008) ($\alpha = 0.35, \beta = 0.65, \gamma = 0$). For TSPR, we use the equation in Haveliwala (2002) to calculate the authority score under given topic in ODP. For TLA, we get the similar result of our approach following Nie et al. (2006).

4.3 Experimental evaluation

We conduct a series of experiments of baseline algorithms and our approach. We got totally five rankings, two of them without topic by PageRank and FolkRank, three of them with topic by TSPR,TLA and TSTR.

To evaluate the proposed approach, we collect top 100 resources for each popular tag to compare rankings with the baselines. In total, 2500 rankings are obtained. Table 2 illustrates several exemplary results. We can see the tag ranking lists are better than the original ones in terms of tag relevance. For instance, both Resource 2 and 3 are tagged with the tag *architecture*. The tag *architecture* ranks at the top position, because it is the most relevant tag in the tag list of Resource 2, but it locates at the second position in the tag list of Resource 3 since it is less important than the tag *web*.

Table 2 Tag ranking examples

No.	URL	Description	Original tags*	Ranked tags
1	http://webdesignfromscratch.com/	London web designers Scratch media London web design agency	Design, inspiration, blog, tutorials, web design, free	Webdesign, design, tutorials, inspiration, blog, free
2	http://twistedsifter.com/2009/12/oshatz-wilkinson-tree-house	Canopy living: the ultimate tree house|TwistedSifter	House, ecology, design, treehouse, architecture	Architecure, hourse, design, ecology, treehouse
3	http://hinchcliffe.org/archive/2009/12/14/18179.aspx	A Web-Oriented Architecture (WOA) Un-Manifesto	web2.0, architecture, woa, rest, principles, web	web, architecture, web2.0, woa, principles, rest
4	http://www.flickr.com	Welcome to Flickr - Photo Sharing	Flickr, photograph, photo, sharing	Photo, share, Flickr, photogragh

* Original tags are from Delicious.com

Table 3 Kendall's τ of five rankings*

	PageRank	FolkRank	TSPR	TLA	TSTR
PageRank	1.000	–	–	–	–
FolkRank	0.726	1.000	–	–	–
TSPR	0.566	0.578	1.000	–	–
TLA	0.612	0.593	0.794	1.000	–
TSTR	0.597	0.526	0.810	0.843	1.000

* Significant at 5% level

We use two performance evaluation measures. First, we use the Kendall's correlation τ in Haveliwala (2002) to measure the degree to which the relative orderings of the top n tag of two rankings are in agreement. The bigger the τ value is,the more harmonious the two rankings will have. In Table 3, the τ value of PageRank and FolkRank is relatively larger than the value of PageRank and other approaches, respectively. This indicates PageRank and FolkRank are more harmonious and nearly rank in the same order. Meanwhile, we can see TSPR, TLA and TSTR are pairwise significant positive correlation. The τ value between them is relative larger. This may be attributed to the topicality of TSPR, TLA and TSTR.

The second measure is Normalized Discounted Cumulated Gain (NDCG) (Jarvelin and Kekalainen (2002)), which is a measure of cumulated gain-based evaluation of information retrieval techniques. We invite 22 students to label each tag of every ranking as one of four levels of relevance: (0) irrelevant; (1) marginally relevant; (2) fairly relevant; (3) highly relevant. In general, the average level represents high agreement on the ranking quality. We calculate the average NDCG as (Liu et al. 2009):

$$N_n = Z_n \sum_{i=1}^{n} (2^{r(i)} - 1)/log_2(1+i) \tag{4}$$

where $r(i)$ is the relevance level of the ith tag and Z_n is a normalization constant.

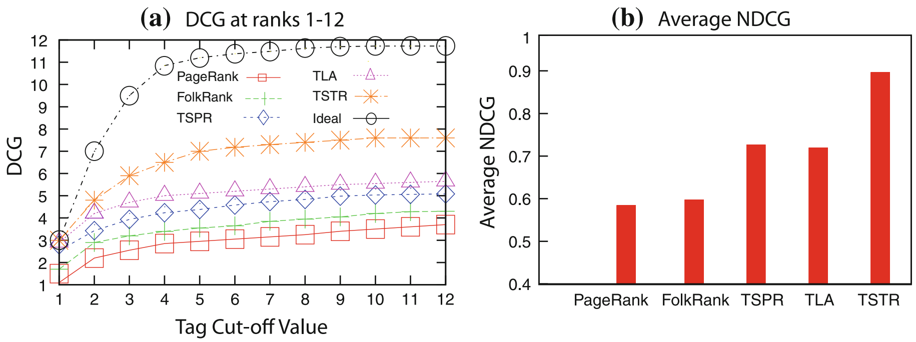

Fig. 2 Performance comparison of different tag ranking strategies

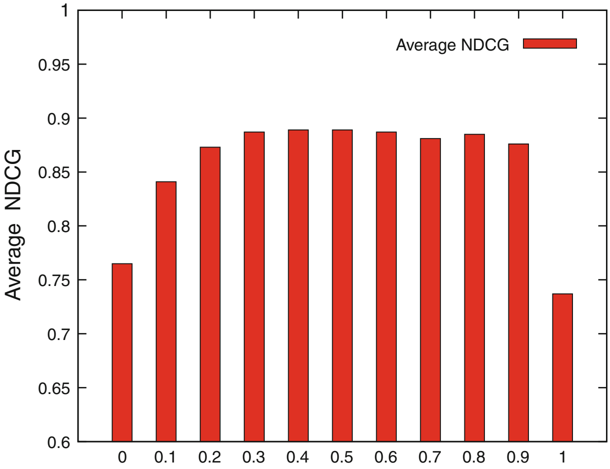

Fig. 3 The performance curve of our tag ranking with respect to the parameter α

Before computing the NDCG, we calculate the Discounted Cumulated Gain (DCG) (Jarvelin and Kekalainen 2002) for each algorithm at different depths as shown in Fig. 2a. The ideal curve becomes nearly a horizon line at 4, which indicates that all relevant tags have been found at 4 practically. The best (TSTR) hangs below the ideal by 1-4 points (9–33)%. The others remain further below by 2-8 points (18–60)%. This indicates TSTR can find relevant tags faster than others. Finally, we calculate the NDCG, shown in Fig. 2b. We can see that TSPR, TLA and TSTR can bring better order to tags. This can be subscribed to the topicality of them. TSTR performs much better than TSPR and TLA (+18%) because topics extracted by LDA are more conformed to the actual situation than the top level of category of ODP.

4.4 Parameter tuning

We further conduct experiments to analyze the sensitivity of our tag ranking scheme with respect to the two parameters (see Eq. 3): α and β. The parameter α controls the impact of probabilistic relevance scores in the random walk process. The parameter β is a transition

Table 4 A subset of urls co-occured on delicious.com and ODP website and their tags

URLs	Original tags	Recommended top 5 tags	Category in ODP
http://photobucket.com/	Images, sharing, photography, useful, billedbehandling, hosting, photos, entrepreneur, visuals, photobucket, Fotografa_Recursos, Redes_sociales, photosharing, pictures, mcshep, manip, hot, diaporama, foto, flickr, imported, album	Photo, photograph, images, photobucket, hosting	Computers/Internet /On_the_Web/Web_ App_lications/Photo_Sharing
http://twitterfeed.com/	Private, twitter, Microblogging, Integration, twitter feed, microblogging, hellotxt, identi.ca, hellotxt, rss_feed, twitterapp, identi.ca, updates, news-reader, social-applications, social-web, news-aggregator, life-stream, feed-stream, socail-apps, feed-reader, google-reader, news-feed, forward, third, rss, feed, mashup, web2.0, syndi	Web2.0, feed, rss, social, mashup	Computers/Internet /On_the_Web/Online _Comm_unities/Social _Network_ing/Twitter
http://www.linkedin.com/	Grupo, rede, create_your_own, jobsearch, career, web2.0, carreira, virtual_community, trabajo, communities, jobs, temp, reseaux_sociaux, connections, socialnetwork, work, divyajyoti, linkedin	Web2.0, likedin, social, career, job	Computers/Internet /On_the_Web/Online _Communities/Social_Networking
http://www.cplus-plus.com/doc/ tutorial/	Cpp, cplusplus, c/c++, programming, tutorial, c_language	Cpp, opp, program, c++, language	Computers/Programming /Languages/C++/ FAQs, Help, and Tutorials/General_C++_Introductions
http://www.clopinet.com/isabelle/Projects/ SVM/applist.html	Machine-learning, datamining, prediction, svm, ai, machine, cs	Svm, machine, machine-learning, ai, learning	Computers/Artificial_ Intelligence /Support_Vector_Machines

Table 5 Performance of tag recommendation

	P@1	P@5	P@10
Delicious.com	6275	0.6051	0.5393
TSPR	0.7858	0.7967	0.6932
TLA	0.8325	0.7189	0.6158
TSTR	0.8975	0.8689	0.8145

probability from z_i to z_j to keep topic continuity. First we follow traditional experience and simply set α to 0.15. For β, we range it from 0 to 1.0. Figure 3 illustrates the results. From the results we can see the performance curve is smooth when β varies in a wide range $[0.2 \sim 0.8]$. We can also see that the result is always better than PageRank, FolkRank, TSPR and TLA when β ranges from 0.1 to 0.9. This indicates the robustness of our approach. According to the results, the optimal values of α and β are around 0.15 and 0.5, respectively.

5 Application

With the emergence of social tagging systems, tag recommendation has become an attractive area of research. Moreover, recommender technologies are usually classified into the following categories, based on how recommendations are made: content-based recommendations, collaborative recommendations and hybrid approaches.

In this paper, we use a hybrid approach to recommend tags for a given resource on delicious.com. TG-HAC (Zeng et al. 2009), a multi-grain hierarchical topic extraction algorithm, was applied to get the topic of given resources. In addition, TSTR approach can be thought as a collaborative recommendation because its intuition is that two tags have the same important under the given resource and topic. Our task is to decide top k tags for a given resource r_i conditional on topic z_k. We followed four steps: (a) find out some URLs, and then downloaded their web pages; (b) use TG-HAC algorithm to get the probability of most general topic of a given resources, $\mu_{r_i z_k}$; (c) calculate $p(t_i|r_i, z_k) = \mu_{r_i z_k} * \phi_{t_i z_k}$; (d) select top k tags. To demonstrate the effectiveness of the tag recommendation method, we randomly choose 100 URLs (a subset shown in Table 4) from our test datasets aforementioned to perform tag recommendation. We use precision as the performance evaluation measure, compared with delicious recommendation algorithm, TSPR and TLA. The average precision with different depths is illustrated in Table 5. From the table, we can see that our tag recommendation results are surprisingly good. The precision is even higher than that of the tags recommended by delicious website and other methods.

6 Conclusion and future work

In this paper we investigate topic based tag ranking problem by using topic-level random walk. We propose a three-step approach for solving this problem. We employ a probabilistic topic model to automatically extract topics from these tags and perform random walk at the topic level on topic tag transition graph to get the authority score for a tag under specific topic. Empirical study shows the importance of using topic of tag with random walk. Finally, we apply purposed approach on tag recommendation. It is proved to be effective even in comparison to delicious website.We assume that each tag has a same topic vector. This is not realistic. In fact, the topic number of tags and topic distribution should vary. We expect to use China Restaurant Problem to solve it in the future.

In this study we investigate the problem about topic-sensitive tag ranking by using topic-level random walk. We propose a three-step approach to rank tags according to their relevance levels. Experimental results show that the proposed approach outperforms the baseline methods, and it is essential to boost the performance of the social-tagging related applications.

Acknowledgments This work is supported by National Natural Science Foundation of China under Grant 60873225, 60773191, 70771043, National High Technology Research and Development Program of China under Grant 2007AA01Z403, Natural Science Foundation of Hubei Province under Grant 2009CDB298, Wuhan Youth Science and Technology Chenguang Program under Grant 200950431171, Open Foundation of State Key Laboratory of Software Engineering under Grant SKLSE20080718, Innovation Fund of Huazhong University of Science and Technology under Grant 2010MS068 and Q2009021, Youth Fund of HBUE (XJ2009013), Humanities and Social Sciences Program of Hubei (2010Q094).

References

Bao S, Wu X, Fei B, Xue G (2004) Folksonomies-cooperative classification and communication through shared metadata http://www.adammathes.com/academic/computer-mediated-communication/folksonomies.html

Bao S, Wu X, Fei B, Xue G, Su Z, Yu Y (2007) Optimizing web search using social annotations. In: Proceedings of WWW2007, pp 501–510

Bao S, Yuan X, Xue G (2007) Optimizing web search using social annotations. In: Proceedings of the 16th international conference on world wide web (WWW2007). ACM, Alberta, pp 501–510

Blei DM, Ng AY (2003) Latent dirichlet allocation. J Mach Learn Res 3:993–1022

Brin S, Lawrence P (1998) The anatomy of a large-scale hypertextual web search engine. Comput Netw ISDN Syst 30:107–117

Guan Z, Wang C, Bu J, Chen C, Yang K, Cai D, He X (2010) Document recommendation in social tagging services. In: WWW '10: Proceedings of the 19th international conference on World wide web (2010), pp 391–400

Haveliwala TH (2002) Topic-sensitive PageRank. In: Proceedings of the 11th international conference on World Wide Web(WWW2002). ACM, Hawaii, pp 517–526

Hotho A, Jaschke R et al. (2006) Trend detection in folksonomies. Lecture notes in computer science 4306:56–70

Hotho A, Jaschke R, Schmitz Ch (2008) Information retrieval in folksonomies: search and ranking, ESWC 2006. Springer, Berlin, 411–426

Jarvelin K, Kekalainen J (2002) Cumulated gain-based evaluation of IR techniques. ACM Trans Inf Syst 20:422–446

Korner C, Benz D, Hotho A, Strohmaier M, Stumme G (2010, Apr 26–30) Stop thinking, start tagging: tag semantics emerge from collaborative verbosity. In: Proceedings of 19th international world wide web conference (WWW2010), Raleigh

Liu D, Hua XS, Yang L (2009) Tag ranking. In: Proceedings of the 18th international conference on World Wide Web (WWW2009). ACM, Madrid, pp 351–360

Nie L, Brian DD, Qi X (2006) Topical link analysis for web search. In: Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval(SIGIR2006). ACM, Washington, pp 91–98

Tang J, Leung H-f, Luo Q, Chen D, Gong J (2009) Towards ontology learning from folksonomies. In: IJCAI09, pp 2089–2094

Tony H, Timo H, Ben L, Joanna S (2005) Social bookmarking tools (I), a general review. D-Lib Magazine, 11(4)

Yang Z, Tang J, Zhang J, Li J (2009) Topic-level random walk through probabilistic model. In: Proceedings of the joint international conferences on advances in data and web management, Springer, pp 162–173

Zeng J et al (2009) Multi-gain hierarchical topic extraction algorithm for text mining. Expert Syst Appl 10:221–232