# Topic-sensitive Tag Ranking

Yan'an Jin[1,2], Ruixuan Li[1]*, Zhengding Lu[1], Kunmei Wen[1], Xiwu Gu[1]
[1]*College of Computer Science and Technology, Huazhong University of Science and Technology, China*
[2]*College of Information and Management, Hubei University of Economics, China*
*jin.yan.an@smail.hust.edu.cn,{rxli,zdlu,kmwen,guxiwu}@hust.edu.cn*

## 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. In this paper, we propose a topic-sensitive tag ranking (TSTR) approach to rate the tags on the web. We employ a generative probabilistic model to associate each tag with a distribution of topics. Then we construct a tag graph according to the co-tag relationships and perform a topic-level random walk over the graph to suggest a ranking score for each tag at different topics. Experimental results validate the effectiveness of the proposed tag ranking approach.*

## 1. Introduction

Social tags have recently emerged as a popular way to allow users to annotate and categorize web resources by assigning one or more descriptive words (called tags). The assigned tags can benefit many applications. For example, Bao et al. [2] utilize tags to optimize web search and Tang et al. [12] 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.

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. The importance of the tag with respect to two different meanings (topics) would be quite different. When a user searches for a document on Delicious.com and his intention is to find some information about Tiger Woods, the tag "*tiger*" would be very useful to help identify the information. However, if the search intention is about a creature, the tag "*tiger*" will be not so useful. How to differentiate the importance of a tag on different topics is thus a challenging problem.

Previously, quite a few works have been conducted to measure the importance of web page. PageRank [11] is one of the state-of-the-art algorithms for this purpose. Haveliwala et al. [5] and Nie et al. [10] further extend the algorithm by calculating a vector of scores to distinguish the importance on different topics. However, as suggested by Hotho et al. [7], PageRank cannot be applied directly on folksonomy because of its two characters: short snippets and undirected triadic hyperedges. Liu et al. [9] propose a ranking scheme to rank tags of a given image on Flickr according to their relevance to the image content. However, the method does not consider the topic information.

In this paper, we propose a topic-sensitive tag ranking (TSTR) approach. In particular, we first employ a generative probabilistic model to associate each tag with a distribution of topics. Then we construct a tag graph according to the co-tag relationships and perform a topic-level random walk over the tag graph to assign importance scores to each tag at different topics. Experiments show that the tag ranking method clearly outperforms the baseline methods.

## 2. Tag ranking based on topics

### 2.1. Preliminary

Basically, the input of tag authority is a collection of tagged resources $\mathbf{D}=\{(r_1, \boldsymbol{t}_1),\cdots, (r_n, \boldsymbol{t}_n)\}$, where $r_i\in$

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

1051-4651/10 $26.00 © 2010 IEEE
DOI 10.1109/ICPR.2010.159

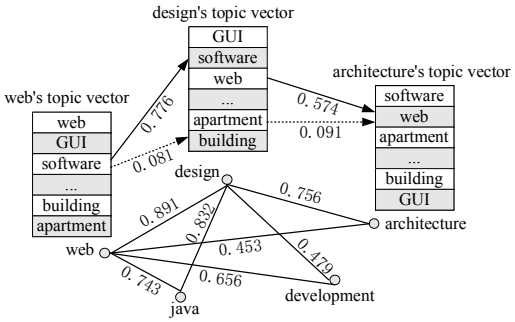


**Figure 1. Illustration of topics within tags.**

**R** is a resource (e.g., document or web page) and $\boldsymbol{t}_i$ is a set of tags assigned to resource $r_i$. We denote **T** as the vocabulary of tags, thus we have $t_{ij} \in \mathbf{T}$. Further we assume that each tag has *k* multiple submeanings (topics). Formally, we denote them as a vector $\boldsymbol{z}$=($z_1$,$z_2$,$\cdots$,$z_k$). Given this, the goal of this work is to rank tags on different topics.

The basic idea of our approach is to incorporate topic distribution into the representation of each tag as well as the importance score , as shown in Figure 1. Therefore, there are two vectors associated with each tag: the topic vector and the authority vector. The topic vector $\theta_t$ is calculated by a widely used topic model, named Latent Dirichlet Allocation (LDA) [3]. Its element $\theta_{t_i z_k}$ represents the relative contribution from topic $z_k$ with tag $t_i$. The authority vector $\varphi_t$ measures the importance of the tag, whose element $\varphi_{t_i z_k}$ denotes tag $t_i$'s importance score on topic $z_k$. This vector is obtained from a topic-level ranking algorithm, which is dynamic during authority propagation. Then the challenge becomes how to determine the topics of the tags and how to calculate the importance of the tags on each topic.

## 2.2. Topic 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 [4]. Gibbs sampling iterates multiple times over each tag $t_i$, and samples a new topic $j$ for the tag based on the conditional probability 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 $\theta_t$. In Table 1, we list three example topics which are extracted out of 100 topics by Gibbs sampling algorithm after 1000 iterations. The extracted topics are *java development*, *social tagging* and *web design*.

**Table 1. Three example topics extracted from the dataset.**

| Topic 5 | | Topic 57 | | Topic 81 | |
|---|---|---|---|---|---|
| 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 |

## 2.3. 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 [7], 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

- $\boldsymbol{V} = \{t_1, ..., t_j\}$ 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: $w_{ij} = \frac{\boldsymbol{N}_{(t_i,t_j)}}{\sum_{k_l}^{|T|} \boldsymbol{N}_{t_k,t_l}}$, where $\mathbf{N}(t_i, t_j)$ is the co-occurrence number of $t_i$ and $t_j$.
- $\Gamma$ 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: $m_{ij} = p(t_j, z_k|t_i, z_l) = \frac{N(t_{ik} \to t_{jl})}{\omega_{ij}}$.

## 2.4. 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 $\alpha$ 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 $\beta$. 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 Figure 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 tag 'design' to tag '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:

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

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:

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

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

$$\begin{aligned} \varphi_{t_j z_l} &= \sum_{i:i\rightarrow j} P(t_j,z_l|t_i,z_l,TS)P(TS|t_i,z_l)\varphi_{t_i z_l} \\ &+ \sum_{i:i\rightarrow j}\sum_{k\in Z} P(t_j,z_l|t_i,z_l,TJ)P(TJ|t_i,z_k)\varphi_{t_i z_k} \\ &= (1-\alpha)\beta \sum_{i:i\rightarrow j} \theta_{t_i z_k} w_{ij} m_{kk} \varphi_{t_i z_l} \\ &+ (1-\alpha)(1-\beta) \sum_{i:i\rightarrow j} \theta_{t_i z_k} w_{ij} \sum_{k\in Z} m_{lk} \varphi_{t_i z_k} \end{aligned} \quad (3)$$

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

| No. | URL | Description | Original tags | Ranked tags |
|---|---|---|---|---|
| 1 | http://webdesignfromscratch.com/ | London web designers Scratch media London web design agency | design,inspiration,blog,tutorials,webdesign,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, architecure,treehouse, design, | 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 |

*Note: Original tags are from Delicious.com.

**Figure 2. Tag ranking examples.**

# 3. Experiments

## 3.1. Experimental settings

All the experiments in this work are conducted on a dataset crawled from Delicious.com. We reduce some triples whose URLs do not match with URL entity on ODP (Open Directory Project, http://www.dmoz.org/) [1]. We preprocess the dataset by a) removing stopwords; b) lower-casing the obtained words, tags, user names; and c) removing tags that appear less than 3 times. Statistically, there are $|T| = 47687$ tags, $|U| = 18273$ users, $|R| = 21011$ resources.

We select some popular tags including *java, design, architecture, web, css* as query keywords to perform tag-based search. To compare, we use PageRank, FolkRank, topic-sensitive PageRank(TSPR) [5], topic link analysis for web search (TLA) [10] and our topic-based random walk algorithm, TSTR to rank the query results,respectively. For consistency, we run them over the same tag graph like Figure 1. The weight $\mathbf{N}(t_i,t_j)$ is set to at least greater than 10. In addition, we fix the number of topics as 100 and run the Gibbs sampling for 1000 iterations each on a 2GHz PC workstation.

## 3.2. Experimental results and evaluation

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. Figure 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*.

We use two performance evaluation measures. First, we use the Kendall's correlation $\tau$ in [5] to measure

**Table 2. Kendall's $\tau$ 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.

the degree to which the relative orderings of the top n tag of two rankings are in agreement. The bigger the $\tau$ value is,the more harmonious the two rankings will have. In Table 2, the $\tau$ 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 TRWA are pairwise significant positive correlation. The $\tau$ value between them is relative larger. This may be attributed to the topicality of TSPR, TLA and TRWA.

The second measure is Normalized Discounted Cumulated Gain (NDCG) [8], 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 [9]:

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

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

Before computing the NDCG, we calculate the Discounted Cumulated Gain(DCG) [8] for each algorithm at different depths as shown in Figure 3(a). 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 Figure 3(b). 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. Conclusion

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

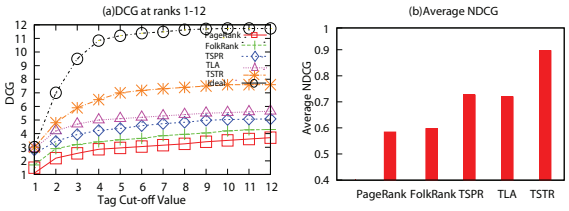


**Figure 3. Performance comparison of different tag ranking strategies.**

that the proposed approach outperforms the baseline methods, and it is essential to boost the performance of the social-tagging related applications.

## 5. Acknowledgements

This work is supported by NSFC (60873225, 60773191, 70771043), 863 Program (2007AA01Z403), NSF of Hubei (2009CDB298), Wuhan Chenguang Program (200950431171), Innovation Fund of HUST (Q2009021), Open Foundation of SKLSE (SKLSE20080718), Youth Fund of HBUE (XJ2009013), Humanities and Social Sciences Program of Hubei (2010Q094).

## References

[1] The open directory project:web directory for over 2.5 million urls. Website. `http://www.dmoz.org`.

[2] S. Bao, X. Wu, B. Fei, G. Xue, Z. Su, and Y. Yu. Optimizing web search using social annotations. *In Proc. of WWW2007*, pages 501–510, 2007.

[3] D. M. Blei, Y. N. Andrew, and M. I. Jordan. Latent dirichlet allocation. *Journal of Machine Learning Research*, 1(3):993–1022, 2003.

[4] T. Griffiths and M. Steyvers. Finding scientific topics. *PNAS*, 101(suppl.1):5228–5235, 2004.

[5] T. H. Haveliwala. Topic-sensitive pagerank. *In Proc. of WWW2002*, pages 517–526, 2002.

[6] P. Heymann and D. Ramage. Social tag prediction. *In Proc. of SIGIR2008*, pages 531–538, 2008.

[7] A. Hotho, R. Jaschke, C. Schmitz, and G. Stumme. Information retrieval in folksonomies: Search and ranking. *In Proc. of ESWC2006*, pages 411–426, 2006.

[8] K. Jarvelin and J. Kekalainen. Cumulated gain-based evaluation of IR techniques. *TOIS*, 4(20):422–446, 2002.

[9] D. Liu, X.-S. Hua, and L. Yang. Tag ranking. *In Proc. of WWW2009*, pages 351–360, 2009.

[10] L. Nie, B. Davison, and X. Qi. Topical link analysis for web search. *In Proc. of SIGIR2006*, pages 91–98, 2006.

[11] L. Page, S. Brin, R. Motwani, and T. Winograd. The pagerank citation ranking: Bringing order to the web. Technical Report SIDL-WP-1999-0120, Stanford University, 1999.

[12] J. Tang, H.-f. Leung, Q. Luo, D. Chen, and J. Gong. Towards ontology learning from folksonomies. In *IJCAI'09*, pages 2089–2094, 2009.