Random Walks on Context-Aware Relation Graphs for Ranking Social Tags

*Han Li** *Zhiyuan Liu** *Maosong Sun*
Department of Computer Science and Technology
State Key Lab on Intelligent Technology and Systems
National Lab for Information Science and Technology
Tsinghua University, Beijing 100084, China
{lihan.thu,lzy.thu}@gmail.com, sms@tsinghua.edu.cn

ABSTRACT
Social tagging provides an efficient way to manage online resources. In order to collect more social tags, many research efforts aim to automatically suggest tags to help users annotate tags. Many content-based methods assume tags are independent and suggest tags one by one independently. Although it makes suggestion easier, the independence assumption does not confirm to reality, and the suggested tags are usually inconsistent and incoherent with each other. To address this problem, we propose to model context-aware relations of tags for suggestion: (1) By regarding resource content as context of tags, we propose Tag Context Model to identify specific context words in resource content for tags. (2) Given a new resource, we build a context-aware relation graph of candidate tags, and propose a random walk algorithm to rank tags for suggestion. Experiment results demonstrate our method outperforms other state-of-the-art methods.

TITLE AND ABSTRACT IN CHINESE

在上下文感知的关系网络中随机游走进行社会标签排序

社会标签是一种有效的管理在线资源的方式。为了获取更多的社会标签，许多研究致力于自动推荐标签来帮助人们标注。很多基于内容的方法认为标签之间是独立的，因此孤立地推荐每个标签。虽然这让推荐方法更简单，但是独立性假设并不符合真实情况，从而导致推荐的标签互相之间存在不一致和不协调。为了解决这一问题，我们提出对标签之间的上下文感知的关系进行建模：（1）我们将资源内容作为标签的上下文，通过标签上下文模型从内容中发现标签的特定上下文。（2）当给定一个新的资源，我们建立候选标签之间的上下文关系图，通过随机游走算法对标签排序并推荐。实验结果证明我们的方法优于已有方法。

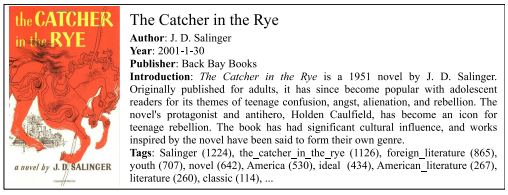
KEYWORDS: tag ranking, context-aware relation, random walk, social tag suggestion.
KEYWORDS IN CHINESE: 标签排序, 上下文感知的关系, 随机游走, 社会标签推荐.

∗ indicates equal contributions from these authors.

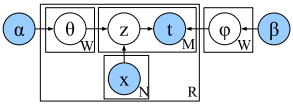
Proceedings of COLING 2012: Posters, pages 653–662,
COLING 2012, Mumbai, December 2012.

1 Introduction

Web 2.0 technologies provide a new scheme, social tagging, for users to collect, manage and share online resources (Gupta et al., 2010). In a social tagging system, each user can freely use any words to annotate resources. Figure 1a shows an exemplary book, *The Catcher in the Rye* from Douban, a review website in China. For the book, many tags have been annotated by thousands of users. For example, the tag "Salinger" is annotated by 1,224 users, which indicates the author J. D. Salinger. The figure also shows some meta-data such as the title, the author and a brief introduction. In this paper, we refer to the meta-data of a resource as *content* and the user-annotated tags as *annotation*.



(a) An example book.



(b) TCM

Figure 1: (a) An example book. (b) Graphical model of TCM.

Social tagging provides a convenient management scheme compared to strict taxonomy in libraries. In order to attract more users to contribute social annotations, many social tagging systems facilitate users through automatic tag suggestion. There are two main approaches: graph-based and content-based. The former approach (Jaschke et al., 2008; Rendle et al., 2009) suggests tags according to users' annotation history, while the latter approach (Si et al., 2009, 2010; Liu et al., 2011) according to resource meta-data. Since graph-based methods often suffer from the cold-start problem when they face new users or resources, content-based methods are usually regarded as an important component in social tagging systems especially in the initial stage. In this paper we focus on the content-based approach.

Many social tagging methods are based on *independence assumption*, which is widely adopted in computational linguistics (Manning and Schutze, 2000) and information retrieval (Manning et al., 2008). Under this assumption, tags are regarded independent with each other given the resource. Although this makes methods easier to implement, it does not accord with the real world, in which the annotated tags of a resource are usually semantically correlated with each other. Hence, if we can find an effective approach to model tag relations, it may improve the suggestion quality significantly.

It is non-trivial to model tag relations. Given a resource, the tag relations are *context-aware*. Two tags may be more related with each other given a resource but less given another one. Moreover, tag relations are complex for modeling. Since tags are not restricted within a pre-defined vocabulary, their relations cannot be well covered by manually-annotated dictionaries such as WordNet (Miller et al., 1990). Hence, we have to statistically learn the semantic relations of tags from a set of annotated data. This will better guarantee the semantic consistency of suggested tags.

To consider the context-aware relations of tags given a resource, tag graphs are a straightforward representation. We consider a random walk method on context-aware tag relation

graphs to rank social tags. There are two critical challenges for this method: (1) How to statistically model the context-aware relations of tags from a large collection of annotation data? (2) After obtaining the context-aware relations of tags, how to construct a tag graph given a resource for random walks? To address the challenges, we propose a probabilistic model to learn the context-aware relations of tags, and propose a random walk algorithm over context-aware relation graphs to suggest tags. To investigate the efficiency of our method, we carry out experiments using real-world datasets.

Related work. Measuring semantic relations have been studied in many tasks such as measuring term similarities (Lin, 1998; Gabrilovich and Markovitch, 2007) and query similarities (Wen et al., 2002; Mei et al., 2008). Meanwhile, context-aware setting is being considered in many applications including recommender systems (Adomavicius and Tuzhilin, 2011) and query suggestion (Brown and Jones, 2001), which is a critical research issue for all applications under real-world complex scene. In social tagging, co-occurrence-based tag relations have been explored to group tags into clusters (Wu et al., 2006b; Brooks and Montanez, 2006; Shepitsen et al., 2008), and have been adopted in personalized tag suggestion (Shepitsen et al., 2008) and extending ontology (Mika, 2005; Wu et al., 2006a). Some specific relations of tags such as subsumption are also studied in social tagging (Si et al., 2010). These relations are mostly context-free. There has been little work on modeling context-aware tag relations for content-based social tag suggestion.

2 Learning Context-Aware Relations of Tags

A resource is denoted as $r \in R$, where R is the set of all resources in the social tagging system. Each resource is composed of the content (meta-data) and the annotation (a set of tags). The content is represented as a sequence of words $\mathbf{x}_r = \{x_i\}_{i=1}^{N_r}$, where N_r is the number of words in $\mathbf{x}_r$. The vocabulary of the words in contents is W, and each word $x_i = w \in W$. The annotation of resource r is represented as $\mathbf{a}_r = \{a_i\}_{i=1}^{M_r}$, where M_r is the number of annotated tags in $\mathbf{a}_r$. The vocabulary of annotations is T, and each annotation $a_i = t \in T$. Tag relations can be either context-free or context-aware, and either symmetric or asymmetric. Without loss of generality, we consider symmetric relations all through the paper and introduce context-free and -aware relations in detail.

2.1 Context-Free Relations

Context-free relations of tags leave context information out. There are various methods to statistically measure context-free relations of tags. The basic idea is regarding two tags are correlated with each other if they tend to be assigned to the same resources. For example, the tags "the_catcher_in_the_rye" and "Salinger" can be found correlated since they are usually assigned to the same resources. In this paper, we measure context-free relations of two tags t_1 and t_2 using joint probability $\Pr(t_1, t_2)$, estimated according to co-occurrences of tags as $\Pr(t_1, t_2) = N_{t_1,t_2}/|R|$, where N_{t_1,t_2} is the number of resources where both t_1 and t_2 appear together, and $|R|$ is the total number of resources.

2.2 Co-Occurrence-based Context-Aware Relations

In this paper, we regard words in resources as the crucial context of tags. The context-aware relation between tags t_1 and t_2 given a context word w can be represented as the conditional probability $\Pr(t_1, t_2|w)$.

We first introduce a naive method to measure context-aware tag relations, i.e., co-occurrence-based context-aware relations. In this method, the conditional probability $\Pr(t_1, t_2|w)$ is estimated according to the co-occurrences of t_1, t_2 and w within a collection of annotated resources as $\Pr(t_1, t_2|w) = N_{w,t_1,t_2}/N_w$, where N_w is the number of resources where w appears, N_{w,t_1,t_2} is the number of resources where w, t_1 and t_2 appear together.

The co-occurrence-based context-aware tag relations are straightforward and easy for implementation. However, empirical experiments show that this type of relations usually suffers from poor performance. The reason is that, in many cases, given two tags of a resource, not all words in the resource can be regarded as their context. It is obvious that each annotated tag usually represents some aspects of a resource, and thus may only correspond to some specific words in the resource.

In order to better model context-aware relations of tags, it is crucial to exactly find corresponding context words for tags. Therefore, we propose a probabilistic graphical model, Tag Context Model (TCM), to learn context words for tags.

2.3 TCM and TCM-based Context-Aware Relations

Tag Context Model (TCM). We propose TCM to find context words of tags. TCM can be regarded as a generative process of each resource r as shown in Figure 1b. Essentially, TCM models semantic relations between words and tags, similar to WTM (Liu et al., 2011) and TAM (Si et al., 2010). We denote the context word sequence as $\mathbf{z}_r = \{z_i\}_1^{M_r}$, corresponding to the tag sequence $\mathbf{a}_r$. The learning goal of TCM is to infer the multinomial distribution of each tag t given word w (i.e., ϕ with $\phi_{tw} = \Pr(t|w)$) and the multinomial distribution of each word w being selected as context word in resource r(i.e., θ with $\theta_{wr} = \Pr(w|r)$). α and β are hyper-parameters of θ and ϕ following Dirichlet distributions.

Given the observed words in resource content $\mathbf{x}$, the joint distribution of θ, ϕ, context words $\mathbf{z}$, and tags $\mathbf{a}$ is $\Pr(\mathbf{z}, \theta, \phi, \mathbf{a}|\mathbf{x}, \alpha, \beta) = \Pr(\theta|\alpha)\prod_{i=1}^{M}\Pr(a_i|z_i, \mathbf{x}, \phi)\Pr(\phi|\beta)\Pr(z_i|\mathbf{x}, \theta)$. The key inference problem of TCM learning is computing posterior distribution of the hidden variables given resource content and tags. The hidden variables in TCM are $\mathbf{z}$, i.e., the context words that correspond to the annotated tags of resources. Here we integrate out the parameters θ and ϕ because it can be regarded as the statistics of the associations between the observed annotations $\mathbf{a}$ and the corresponding $\mathbf{z}$.

In this paper we select Gibbs Sampling for inference, which has been widely adopted in graphical models such as LDA (Griffiths and Steyvers, 2004). Since we integrate out θ and ϕ, the inference algorithm is also referred to as *collapsed* Gibbs Sampling. In Gibbs Sampling, we compute the conditional probability as

$$\Pr(z_i = w|\mathbf{z}_{\neg i}, a_i = t, \mathbf{a}, \mathbf{x}, \alpha, \beta) = \frac{\Pr(\mathbf{z}, \mathbf{a}|\mathbf{x}, \alpha, \beta)}{\Pr(\mathbf{z}_{\neg i}, \mathbf{a}|\mathbf{x}, \alpha, \beta)} \propto \frac{N_{tw}^{\neg i} + \beta}{\sum_t N_{tw}^{\neg i} + |T|\beta} \times \frac{N_{wr}^{\neg i} + \alpha}{\sum_w N_{wr}^{\neg i} + |W|\alpha}, \quad (1)$$

where $N_*^{\neg i}$ indicates the annotation a_i is excluded, N_{tw} is the number of times that tag t takes w as its context word, and N_{wr} is the number of times that word w is selected as a context word within resource r. Note that the probability shown in Equation (1) is unnormalized. The actual probability of assigning a tag to context word w is computed by dividing the quantity in Equation (1) for word w by summing over all unique words in the resource content. Gibbs Sampling outputs the estimation of $\mathbf{z}$ for annotated tags. We

further estimate ϕ and θ as $\phi_{tw} = \Pr(t|w) = \frac{N_{tw}+\beta}{\sum_t N_{tw}+|T|\beta}$ and $\phi_{wr} = \Pr(w|r) = \frac{N_{wr}+\alpha}{\sum_w N_{wr}+|W|\alpha}$. With the estimated ϕ, we can obtain all context words of each tag t, i.e., the words that have higher values of $\phi_{tw} = \Pr(t|w)$. Based on the estimations, we further measure context-aware relations of tags.

TCM-based Context-Aware Relations. We define a function $\delta(x)$ as $\delta(x) = 1$ if x is true, otherwise $\delta(x) = 0$. We calculate TCM-based context-aware relations of two tags t_1 and t_2 using **a** and **z** as follows:

$$\Pr(t_1, t_2|w) = \frac{\sum_{r\in R} \delta_r(z_i = w \cap z_j = w \cap a_i = t_1 \cap a_j = t_2)}{\sum_{r\in R} \delta_r(z_i = w)}. \quad (2)$$

In this equation, if $\delta_r(z_i = w \cap z_j = w \cap a_i = t_1 \cap a_j = t_2) = 1$, it indicates the resource $r \in R$ has two tags t_1 and t_2 and both of them are assigned to w as their context word; if $\delta_r(z_i = w) = 1$, it indicates the resource $r \in R$ has w being assigned as context word.

The TCM-based context-aware relations calculated in Equation (2) have a potential size of $|T|^2|W|$, where $|T|$ is the vocabulary size of tags and $|W|$ is the vocabulary size of words. The estimation of context-aware relations suffers from more serious problem of sparsity compared to context-free relations. To alleviate the sparsity problem, we introduce a remedy solution: linear interpolation smoothing. We use the conditional-independent context-aware relations for interpolation. Suppose two tags t_1 and t_2 are conditionally independent given w, the context-aware relations will be $\Pr^+(t_1, t_2|w) = \Pr(t_1|w)\Pr(t_2|w)$, and the interpolation smoothing is performed as $\Pr^*(t_1, t_2|w) = \lambda \Pr(t_1, t_2|w) + (1-\lambda)\Pr(t_1|w)\Pr(t_2|w)$, where λ is the interpolation factor. In this paper, we simply set $\lambda = 0.5$.

Given a resource r with its content $\mathbf{x}_r$ as context, the context-aware relation of two tags t_1 and t_2 can be calculated according to their context-aware relation given each word in the resource content as context word, $\Pr(t_1, t_2|r) = \Pr(t_1, t_2|\mathbf{x}_r) = \sum_{w\in\mathbf{x}} \Pr(t_1, t_2|w)\Pr(w|\mathbf{x})$.

3 Random Walks for Ranking Tags

After modeling the context-aware relations of tags, we can build a context-aware relation graph of tags and rank tags by random walks over the graph.

3.1 Context-Aware Relation Graph Building

Here we focus on building undirected graphs which correspond to symmetric context-aware relations. For a resource r with its content $\mathbf{x}$, we first rank tags according to the conditional probability of each tag estimated by TCM, i.e., $\Pr(t|r) = \sum_{w\in\mathbf{x}} \Pr(t|w)\Pr(w|\mathbf{x})$, where $\Pr(w|\mathbf{x})$ is the probability of w being selected as context words within resource content $\mathbf{x}$, and $\phi_{tw} = \Pr(t|w)$ is the probability of w working as context word of t. The measure assumes each tag is conditionally independent given the resource and thus can be calculated separately. With $\Pr(t|r)$ we select top-ranked tags as candidate tags, denoted as T_c. The number of candidate tags, $|T_c|$, can be manually pre-defined, which should be much larger than the number of suggested tags M_r, but much smaller than the size of tag vocabulary $|T|$.

With candidate tags T_c, we build the context-aware relation graph of tags. We denote the graph as $G = \{V, E\}$, where V is the set of nodes with each node $v_i = t_i \in T_c$, and E is the set of edges with each edge links two nodes in V, e.g., $e_{ij} = (v_i, v_j)$. In an undirected

graph, e_{ij} indicates the edge between v_i and v_j with $e_{ij} = e_{ji}$. We set the edge weight using symmetric context-aware relation probability, i.e., $e_{ij} = \Pr(t_i, t_j|r)$, which indicates the semantic relatedness between t_i and t_j given r as context. With G, we represent the context-aware relations of candidate tags given the resource within a unified graph. The next step is performing random walks over the graph to rank tags.

3.2 Random Walks over Context-Aware Relation Graphs

We conduct random walks over context-aware tag relation graphs to rank and suggest social tags. Random walks have been widely used in many tasks of computational linguistics and information retrieval, which can take the knowledge of the whole graph together for ranking nodes (Liu et al., 2009, 2010).

The basic idea of random walks is that a node is important if there are other important nodes connecting with it. Given a tag graph, we denote the ranking score of a node v_i at iteration k as $r_k(i)$. The random walk process is formulated as

$$r_{k+1}(j) = \gamma \sum_{v_i \in N(v_j)} \frac{e_{ij}}{\sum_j e_{ij}} r_k(i) + (1-\gamma)\frac{1}{|V|}, \tag{3}$$

where $\sum_j e_{ij}$ is the out-degree of node v_i, γ is the damping factor ranging from 0 to 1, and $|V|$ is the number of nodes in G. In this paper, we follow most work and set $\gamma = 0.85$ (Langville and Meyer, 2004). The random jump probability in Equation (3) can also be set non-uniformly. Suppose we assign larger scores to some nodes, the final ranking scores will prefer these nodes and their neighbors. The new method is referred to as random walks with restart (RWR) (Tong et al., 2006). RWR takes node preferences into consideration during random walks, which can be written as $r_{k+1}(j) = \gamma \sum_{v_i \in N(v_j)} \frac{e_{ij}}{\sum_j e_{ij}} r_k(i) + (1-\gamma)\Pr(j)$, where $\Pr(j)$ is the preference of node v_j. In this paper, we set $\Pr(j) = \Pr(t_j|r)$ estimated by TCM. Note that $\sum_{v_i \in V} \Pr(i) = 1$. For tag suggestion, we simply use RWR scores to rank candidate tags and select top-ranked ones for suggestion. For this task, we denote the random walk method over context-free relation graphs as CFR; the method over co-occurrence context-aware relation graphs as CCR; and the method over TCM-based context-aware relation graphs as TCM.

4 Experiments

In the previous sections, we introduced the framework of suggesting social tags based on context-aware relations of tags given the resource. To investigate the efficiency of our method, in this section, we carry out experiments on real-world datasets.

4.1 Datasets and Experiment Setting

In our experiments, we select two real world datasets for evaluation. In Table 1 we show statistics of these datasets, where $|R|$, $|W|$, $|T|$, $\hat{N}_r$ and $\hat{M}_r$ are the number of resources, the vocabulary of contents, the vocabulary of tags, the average number of words in each resource content and the average number of tags in each resource, respectively. The two datasets, denoted as BOOK and MUSIC, contain book and music descriptions as content respectively, together with their annotated tags. Both of them are crawled from Douban (`www.douban.com`), the largest Chinese product review service.

Data	$\|R\|$	$\|W\|$	$\|T\|$	$\hat{N}_r$	$\hat{M}_r$
BOOK	26,807	82,420	41,199	368.69	8.95
MUSIC	25,785	107,100	31,288	541.13	8.13

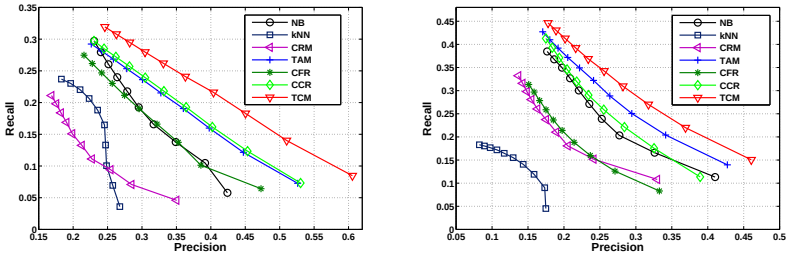
Table 1: Statistical information of two datasets.

We use precision/recall for evaluation. For a resource, we denote gold standard tags as $\mathbf{a}_g$, the suggested tags as $\mathbf{a}_s$, and thus the correctly suggested tags as $\mathbf{a}_g \cap \mathbf{a}_s$. Precision and recall are defined as $P = |\mathbf{a}_g \cap \mathbf{a}_s|/|\mathbf{a}_s|$ and $R = |\mathbf{a}_g \cap \mathbf{a}_s|/|\mathbf{a}_g|$. In experiments, we perform 5-fold cross validation for each method, and the evaluation scores are computed by micro-averaging over resources of test set. We will evaluate the performance when the number of suggested tags M ranges from 1 to 10.

4.2 Evaluation Results

We select Naive Bayes (NB) (Garg and Weber, 2008), kNN (Li et al., 2009), CRM (Iwata et al., 2009) and TAM (Si et al., 2010) as baseline methods for comparison. NB and kNN are representative classification-based methods; while CRM and TAM are representative topic-based methods. We set the parameters of the baselines as follows, by which these methods achieve their best performance: the number of topics $T = 1,024$ for CRM, the number of nearest neighbors $k = 5$. We will also compare three types of tag relations for tag suggestion, i.e., CFR (Section 2.1), CCR (Section 2.2) and TCM (Section 2.3).

In Figure 2a and Figure 2b we show the precision-recall curves of NB, kNN, CRM, TAM, CFR, CCR and TCM on BOOK and MUSIC datasets. Each point of a precision-recall curve represents suggesting different number of tags ranging from $M = 1$ (bottom right, with higher precision and lower recall) to $M = 10$ (upper left, with higher recall but lower precision), respectively. The closer the curve to the upper right, the better the overall performance of the method.



(a) BOOK (b) MUSIC

Figure 2: Precision-recall curves for social tag suggestion on BOOK and MUSIC.

From Figure 2a and Figure 2b, we find that: (1) TCM performs consistently better than other methods for social tag suggestion on both datasets. This indicates the effectiveness and efficiency of TCM. (2) CFR, the method based on context-free tag relations, fails to suggest good tags compared to TCM and some baselines. This indicates the insufficiency of context-free relations and the necessity of modeling context-aware relations for social

tag suggestion. (3) TCM is superior to CCR. It reveals that measuring context-aware relations simply based on co-occurrences between words and tags may introduce many noises because each tag of a resource mostly reflects only some specific words of the resource instead of all of them. This suggests that modeling context-aware tag relations is a non-trivial task, and we have to find corresponding context words for tags so as to build accurate context-aware relations. This is what we do by proposing TCM.

In Table 2, we show top-10 tags suggested by several methods for book *The Catcher in the Rye*, the example in Figure 1a. Here we do not show the results of *k*NN because its performance is too poor to compare with others. The number in the brackets after each method is the count of correctly suggested tags. The correctness of suggested tags are marked with +/−, and the incorrect tags are also highlighted in boldface. From Table 2, we can see that NB, CRM and TAM tend to suggest inconsistent and unrelated tags due to independence assumption, such as "philosophy" and "history". CFR is context-free and its suggested tags are also inconsistent. CCR and TCM take the given resource as context, and thus achieve better performance especially for several top tags. TCM is obviously better than CCR, and can suggest specific tags such as "Salinger" and "the_catcher_in_the_rye".

Method	Suggested Tags
NB(5)	novel (+), foreign_literature (+), literature (+), **history** (-), **philosophy** (-), America (+), classic (+), **China** (-), **Japan** (-), **Chinese_literature** (-)
CRM(5)	novel (+), foreign_literature (+), literature (+), **history** (-), **China** (-), culture (+), **Chinese_literature** (-), classic (+), **Britain** (-), **philosophy** (-)
TAM(5)	novel (+), foreign_literature (+), literature (+), America (+), **Britain** (-), **Chinese_literature** (-), **China** (-), **history** (-), classic (+), **British_literature** (-)
CFR(5)	novel (+), literature (+), foreign_literature (+), **China** (-), **Chinese_literature** (-), classic (+), America (+), **history** (-), **love** (-), **Britain** (-)
CCR(6)	novel (+), foreign_literature (+), literature (+), America (+), classic (+), **Britain** (-), American_literature (+), **Chinese_literature** (-), **China** (-), **Britain_literature** (-)
TCM(10)	novel (+), foreign_literature (+), Salinger (+), literature (+), the_catcher_in_the_rye (+), America (+), American_literature (+), foreign_novel (+), classic (+), youth (+)

Table 2: Suggested tags for book *The Catcher in the Rye* (example in Figure 1a).

Conclusion and Future Work

In this paper, we propose TCM to find context words for tags from resource content. We model TCM-based context-aware tag relations, build a context-aware relation tag graph, and perform random walks over the graph to rank tags. Experiment results show that our method can sufficiently suggest more consistent tags compared to other methods.

We have several research plans: (1) Build a unified method to simultaneously find context words of tags and model context-aware tag relations. (2) Incorporate more context, such as time-stamps and geographical information of annotation. (3) Model context-aware tag relations for other applications to investigate their effectiveness, such as personalized information retrieval and recommender systems.

Acknowledgments

This work is supported by the National Natural Science Foundation of China (NSFC) under the grant No. 61170196 and 61202140. The authors would like to thank Douban for providing data.

References

Adomavicius, G. and Tuzhilin, A. (2011). Context-aware recommender systems. *Recommender Systems Handbook*, pages 217–253.

Brooks, C. and Montanez, N. (2006). Improved annotation of the blogosphere via autotagging and hierarchical clustering. In *Proceedings of WWW*, pages 625–632. ACM.

Brown, P. and Jones, G. (2001). Context-aware retrieval: Exploring a new environment for information retrieval and information filtering. *Personal and Ubiquitous Computing*, 5(4):253–263.

Gabrilovich, E. and Markovitch, S. (2007). Computing semantic relatedness using wikipedia-based explicit semantic analysis. In *Proceedings of IJCAI*, pages 6–12.

Garg, N. and Weber, I. (2008). Personalized, interactive tag recommendation for flickr. In *Proceedings of RecSys*, pages 67–74.

Griffiths, T. and Steyvers, M. (2004). Finding scientific topics. *PNAS*, 101(Suppl 1):5228.

Gupta, M., Li, R., Yin, Z., and Han, J. (2010). Survey on social tagging techniques. *ACM SIGKDD Explorations Newsletter*, 12(1):58–72.

Iwata, T., Yamada, T., and Ueda, N. (2009). Modeling social annotation data with content relevance using a topic model. In *Proceedings of NIPS*, pages 835–843.

Jaschke, R., Marinho, L., Hotho, A., Schmidt-Thieme, L., and Stumme, G. (2008). Tag recommendations in social bookmarking systems. *AI Communications*, 21(4):231–247.

Langville, A. and Meyer, C. (2004). Deeper inside pagerank. *Internet Mathematics*, 1(3):335–380.

Li, X., Snoek, C., and Worring, M. (2009). Learning social tag relevance by neighbor voting. *IEEE Transactions on Multimedia*, 11(7):1310–1322.

Lin, D. (1998). Automatic retrieval and clustering of similar words. In *Proceedings of COLING*, pages 768–774.

Liu, D., Hua, X., Yang, L., Wang, M., and Zhang, H. (2009). Tag ranking. In *Proceedings of WWW*, pages 351–360.

Liu, Z., Chen, X., and Sun, M. (2011). A simple word trigger method for social tag suggestion. In *Proceedings of EMNLP*, pages 1577–1588. Association for Computational Linguistics.

Liu, Z., Huang, W., Zheng, Y., and Sun, M. (2010). Automatic keyphrase extraction via topic decomposition. In *Proceedings of EMNLP*, pages 366–376.

Manning, C., Raghavan, P., and Schtze, H. (2008). *Introduction to information retrieval*. Cambridge University Press New York, NY, USA.

Manning, C. and Schutze, H. (2000). *Foundations of statistical natural language processing*. MIT Press.

Mei, Q., Zhou, D., and Church, K. (2008). Query suggestion using hitting time. In *Proceedings of the 17th ACM conference on Information and knowledge management*, pages 469–478. ACM.

Mika, P. (2005). Ontologies are us: A unified model of social networks and semantics. *Proceedings of ISWC*, pages 522–536.

Miller, G., Beckwith, R., Fellbaum, C., Gross, D., and Miller, K. (1990). WordNet: An on-line lexical database. *International Journal of Lexicography*, 3:235–244.

Rendle, S., Balby Marinho, L., Nanopoulos, A., and Schmidt-Thieme, L. (2009). Learning optimal ranking with tensor factorization for tag recommendation. In *Proceedings of KDD*, pages 727–736.

Shepitsen, A., Gemmell, J., Mobasher, B., and Burke, R. (2008). Personalized recommendation in social tagging systems using hierarchical clustering. In *Proceedings of RecSys*, pages 259–266.

Si, X., Liu, Z., Li, P., Jiang, Q., and Sun, M. (2009). Content-based and graph-based tag suggestion. *ECML PKDD Discovery Challenge 2009*, page 243.

Si, X., Liu, Z., and Sun, M. (2010). Modeling social annotations via latent reason identification. *IEEE Intelligent Systems*, 25(6):42–49.

Tong, H., Faloutsos, C., and Pan, J.-Y. (2006). Fast random walk with restart and its applications. In *Proceedings of ICDM*, pages 613–622.

Wen, J., Nie, J., and Zhang, H. (2002). Query clustering using user logs. *ACM Transactions on Information Systems*, 20(1):59–81.

Wu, H., Zubair, M., and Maly, K. (2006a). Harvesting social knowledge from folksonomies. In *Proceedings of the seventeenth conference on Hypertext and hypermedia*, pages 111–114. ACM.

Wu, X., Zhang, L., and Yu, Y. (2006b). Exploring social annotations for the semantic web. In *Proceedings of the 15th international conference on World Wide Web*, pages 417–426. ACM.