J Intell Inf Syst (2013) 40:307–325
DOI 10.1007/s10844-012-0227-2

Folksonomy link prediction based on a tripartite graph for tag recommendation

Majdi Rawashdeh · Heung-Nam Kim · Jihad Mohamad Alja'am · Abdulmotaleb El Saddik

Received: 14 June 2012 / Revised: 23 October 2012 / Accepted: 24 October 2012 /
Published online: 14 November 2012

Abstract Nowadays social tagging has become a popular way to annotate, search, navigate and discover online resources, in turn leading to the sheer amount of user-generated metadata. This paper addresses the problem of recommending suitable tags during folksonomy development from a graph-based perspective. The proposed approach adapts the Katz measure, a path-ensemble based proximity measure, for the use in social tagging systems. We model a folksonomy as a weighted, undirected tripartite graph. We then apply the Katz measure to this graph, and exploit it to provide tag recommendations for individual users. We evaluate our method on two real-world folksonomies collected from CiteULike and Last.fm. The experimental results demonstrate that the proposed method improves the recommendation performance and is effective for both active taggers and cold-start taggers compared to existing algorithms.

Keywords Folksonomy · Graph-based ranking · Link prediction · Social tagging · Tag recommendation · Tripartite graph

1 Introduction

With the recent prevalence of social (or collaborative) tagging services, such as Last.fm, Delicious, Flicker, and CiteULike, tagging services have become powerful applications for Web users and have drawn more attention from both the industry and the research community. A common feature of many social tagging applications

M. Rawashdeh · H.-N. Kim (✉) · A. El Saddik
School of Electrical Engineering and Computer Science, University of Ottawa,
800 King Edward, Ottawa, ON, Canada, K1N 6N5
e-mail: nami4596@gmail.com

J. M. Alja'am
Department of Computer Science and Engineering, Qatar University, Doha, Qatar

is that they allow their users to assign freely chosen keywords, so-called *tags*, to items (Hotho et al. 2006). Users of social tagging service are flexible and not restricted any more to a rigid hierarchy of content similar to taxonomies or predefined dictionaries. Therefore, social tagging has become a popular way to categorize, share and organize social media content (e.g., videos, images, music, etc), in turn leading to the sheer amount of user-generated metadata and folksonomies (Bischoff et al. 2008).

With this popularity, tag recommendations have attracted increasing recent attention to fully utilize the social tagging systems. Tag recommender systems suggest relevant tags to users during their annotation process in order to help users organize their items. Recent studies tackled the issue of utilizing folksonomies to make personalized tag recommendations from a graph-based perspective. Some of these studies viewed a folksonomy as a tripartite hypergraph in which each hyperedge connects a user, a tag and an item (Hotho et al. 2006; Ramezani 2011). Since folksonomies grow and change rapidly, the need for an approach to predicting accurately the links in the folksonomy graph-based representation becomes an important issue to be addressed. As such, this paper presents a graph-based ranking algorithm that is designed for personalized tag recommendations. Given the fact that a folksonomy can be represented by a tripartite graph as presented in earlier work (Hotho et al. 2006; Ramezani 2011), we formalize the tag recommendation problem as the link-prediction problem. On this graph, we exploit the Katz measure, which is a path-ensemble based proximity measure (Katz 1953), to quantify the proximity of two nodes based on weighted sums over collections of possible paths connecting those nodes. The basic premise behind our algorithm is that appropriate nodes for a given node would be in close proximity to that node from a graph viewpoint. In this context, for recommending tags suited to a given user-item pair, our method speculates as to how a certain node is close not only to that user, but also to that item. Based on the proximity of nodes, we uncover triangle graphs that are likely to appear in the tripartite folksonomy graph.

The remainder of this paper is organized as follows: Section 2 briefly reviews studies related to tag recommendation. In Section 3, we provide basic notations and formalizations. We then describe our model and provide a detailed description of how this model is applied to tag recommendations in Section 4. In Section 5, we present the effectiveness of our method through experimental evaluations. Finally, we give our conclusions in Section 6.

2 Related work

Recent years have seen an increasing number of studies in the area of social tagging services. Some early work in using social tagging is presented by Hotho et al. (2006). The authors proposed a formal model for folksonomies, called FolkRank. It computes a PageRank vector from the tripartite graph induced by the Folksonomy. FolkRank used a folksonomy structure for tag recommendation and ranking search requested within tagging systems. Recently, a User-centric Tag Model (UCTM) is proposed by Wetzker et al. (2010). The algorithm uses a 3-order tensor to model the association between users, items, and tags. The algorithm maps individual tag vocabularies, called personomies, on the corresponding folksonomies with the tagged items. The UCTM can be applied for tag-based search and tag recommendation. Xu

et al. (2008) presented a personalized search framework for item recommendation. This framework has the ability to rank a webpage based on topic matching to the user's interests and the input query in the tag space. A new tag ranking scheme is proposed by Liu et al. (2009). The scheme captures the tag's relevance level to the image content and ranks these tags accordingly. The relevance score of a tag is quantified based on probability density estimation. Thereafter, a random walk over a tag similarity graph is performed to refine the relevance score. This scheme can be applied in tag-based image search, tag recommendation, and image group recommendation.

In recent work, Ramezani (2011) proposed a weighted directed graph which models the informational channels of a folksonomy and applied the PageRank algorithm to this graph so as to enhance graph-based tag recommendation techniques. A modified form of K-Nearest Neighbor for tag recommendation in folksonomies is presented in Gemmell et al. (2009), by incorporating user, resource and tag information into the algorithm. Gemmell et al. (2011) also proposed a linear-weighted hybrid algorithm for tag-based resource recommendation. The algorithm was analyzed across six real-world datasets with different characteristics. In Budura et al. (2009), a tag recommendation method based on the neighborhood of a document-tag graph is proposed. The tag rank relies on the occurrence of the tag in the neighbors of active resource, co-occurrence of the tag to a set of tags inferred from active resource, distance between resources and between tags. Guan et al. (2009) studied personalized tag recommendation by considering association between resources and association between resources to tags. The study considers the preference vector of tags that the user more frequently used to generate personalized tag recommendations. Tags are ranked by incorporating document ranking and active users' most frequent tags in a ranking function.

Lipczak and Milios (2010) proposed a system that recommends tags based on merged scores of user profile related tags and resource related tags. The system uses the content of each added post to update all stored information so that new recommendations reflect users' current interests. Song et al. (2011) tackled the issue of tag recommendation from a machine learning perspective. The authors proposed two document-centered approaches for recommending tags in social networking systems. The first approach is a graph-based approach where the tagged data is represented by two bipartite graphs, and the second one is a prototype-based approach that uses a sparse multiclass Gaussian process classifier for efficient document classification. Rendle and Schmidt-Thieme (2010) presented a new factorization model for tag recommendation which is a special case of the Tucker decomposition. The new factorization model extends the Bayesian personalized optimization criterion to the task of tag recommendation and explicitly reflects the pairwise interactions between users, items and tags. More recently, Hamouda and Wanas (2011) proposed a personalized tag recommendation system for social bookmarking systems using collaborative filtering. The proposed system recommends tags based on similar users and similar bookmarks, also addresses the two limitations of collaborative filtering, first-time seen bookmark that have not been tagged before and cold start users with no sufficient history for recommendation.

Social tagging has also attracted attention in the field of music applications. Bu et al. (2010) proposed a hypergraph model which combines social media information and music acoustic-based content. Another model was proposed by

Horsburgh et al. (2011). They incorporated content-based representation into a tag-based recommender system. This model constructs a tag-track matrix by including audio content into a tag space and then learning hybrid concepts using latent semantic analysis. In another work, Levy and Sandler (2009) built a representation matrix that combines clustered content representation and tags, and then employed probabilistic latent semantic analysis (PLSA) to learn new concepts which generalize both content and tags. Miotto and Orio (2012) presented an approach to music search and discovery based on a graph representation. This approach combines acoustic similarities and tags in a single probabilistic framework. Tags and the acoustic similarity were used together at the same time for ranking music. Font et al. (2012) utilized an online audio sharing site, Freesound, to evaluate the performance of four variants of algorithms on tag recommendation based on the tag semantic similarity derived from tag co-occurrences in the Freesound folksonomy. Symeonidis et al. (2008) presented a tensor model to recommend music according to users' multimodal perception of music, by applying the latent semantic analysis and dimensionality reduction using the higher order singular value decomposition. Tatli and Birturk (2011) proposed a method for creating music recommendation based on the user-supplied tags that are augmented with a hierarchical structure extracted for top level genres from DBpedia. This approach aimed to represent individual tracks (songs) in a lower dimensional space and to use multi-domain information in recommendations. These music-related studies differ from our work in that they attempted to analyze music content, e.g., acoustic or sound features. Therefore, they could not be directly comparable to our work, as our aims were to analyze tripartite relations inherent in a folksonomy and thus to suggest relevant tags to users during their tag annotation process.

3 Definitions and notations

We start by introducing definition and notations exploited in this paper. In social tagging system, users annotate items with tags, creating ternary associations between users, tags, and items. For a set of users $U = \{u_1, u_2, \ldots, u_{|U|}\}$, a set of tags $T = \{t_1, t_2, \ldots, t_{|T|}\}$, and a set of items $I = \{i_1, i_2, \ldots, i_{|I|}\}$, a folksonomy can be defined as a tuple $F := (U, T, I, Y)$ where $Y \subseteq U \times T \times I$ is a ternary relation, called tag assignments (Hotho et al. 2006). From the tag assignments, three matrices can be obtained by aggregating over items, tags, and users, respectively: a $|U| \times |T|$ user-tag matrix $\mathbf{M}_{UT}$, a $|U| \times |I|$ user-item matrix $\mathbf{M}_{UI}$, and a $|T| \times |I|$ tag-item matrix $\mathbf{M}_{TI}$. Each entry of the matrices represents the number of times that the corresponding row and column co-occurred in Y (Wetzker et al. 2010). By using these three matrices, a folksonomy can be converted into an undirected tripartite graph $G = (U \cup T \cup I, E)$ whose nodes can be partitioned into three disjoint sets, U, T, and I, such that every node of each set is adjacent to at least one node in each of the two other sets (Hotho et al. 2006). Each edge (or link) in E connects the corresponding row and column of nonzero entries in the matrices, $\mathbf{M}_{UT}$, $\mathbf{M}_{UI}$, and $\mathbf{M}_{TI}$. In addition, each link has a weight which is equal to an entry's value in the corresponding matrix. Figure 1 shows an example of a weighted undirected tripartite graph that represents a folksonomy.

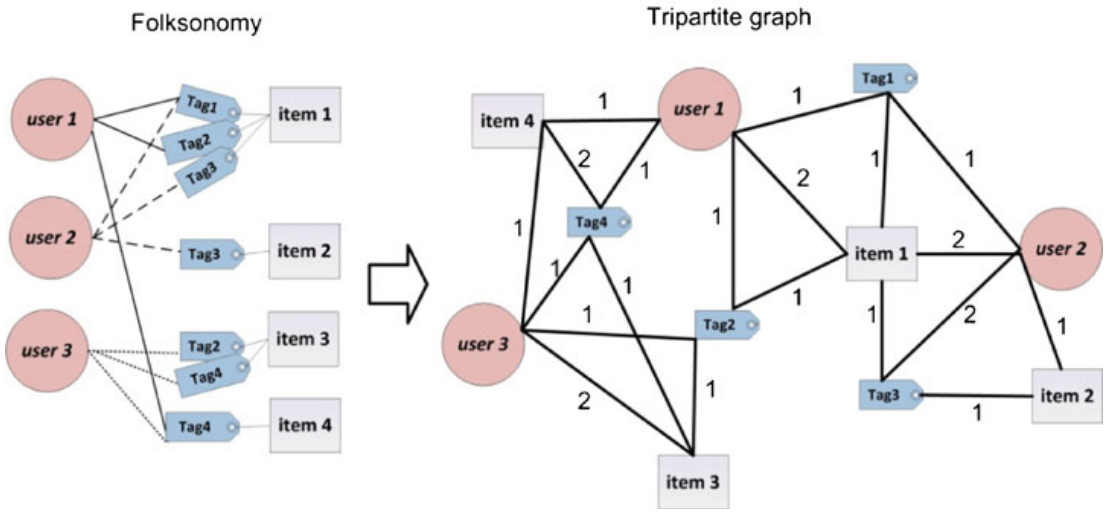

Fig. 1 Transforming a folksonomy to a weighted undirected tripartite graph

Given the fact that a folksonomy can be converted into the undirected tripartite graph G, the problem of our study can be seen as the link-prediction problem for tag recommendation. In general, the goal of personalized tag recommender algorithms is to identify a set of tags suited to a given user-item pair (u, i). From a tripartite graph point of view (as shown in Fig. 2), if user u is interested in tagging item i, then new link (u, i) between user u and item i appear in the graph. To recommend tags for a given link (u, i), our method attempts to predict both the link (u, t) between user u and tag t, and the link (i, t) between item i and tag t. Thereafter, our method calculates a ranking score of tag t based on such predicted links and thus generates a list of top N ranked tags suited to user u for item i.

In social tagging systems, our tag recommendation problem is formally defined as follows: Given a folksonomy graph $G = (U \cup T \cup I,\ E)$, a positive integer N, and a pair of a user node $u \in U$ and an item node $i \in I$, identify a totally ordered set $T(u, i)$ of tag nodes that are likely to appear with the pair (u, i) in the form of a triangle graph such that $|T(u, i)| \leq N$ and $T(u, i) \subseteq T$.

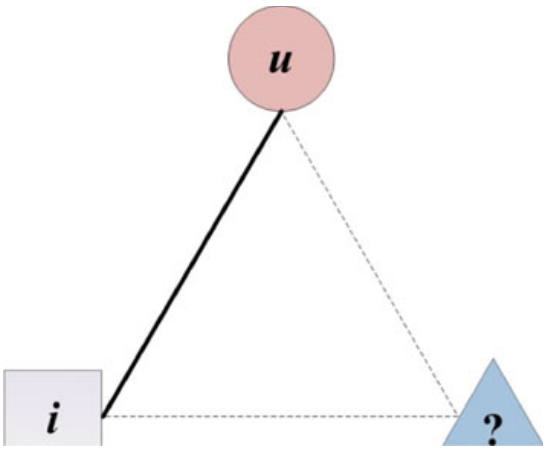

Fig. 2 Triangle link prediction for tag recommendation

4 Folksonomy link prediction for personalized tag recommendation

4.1 Adjacency matrix for folksonomy graph

We begin by defining the adjacency matrix **A** corresponding to the undirected tripartite graph G, as shown in Fig. 3. Each entry in the adjacency matrix is exactly equal to the weight associated with the link between the corresponding two nodes in the graph. Different weighting schemes can be used for assigning weights on links in the graph, which result in different adjacency matrices. In our study we have tested the following weights:

- *Binary*: the weight is either 0 or 1. For example, if user u assigned tag t to item i, then we set the associated weight to the corresponding links (u, t), (u, i), and (t, i) to 1, and 0 otherwise.
- *Frequency*: the weight is similar to *term frequency* used in information retrieval. Here, the weight associated with the links (u, t), (u, i), and (t, i) denotes the number of items user u tagged with tag t, the number of tags user u assigned to item i, and the number of users who tagged item i with tag t, respectively.
- *Okapi BM25*: to increase/decrease the importance of tags within/among users and items, we use a BM25 weighing scheme, as has also been attempted in previous studies (Xu et al. 2008; Vallet et al. 2010). Unlike the TF-IDF weighing scheme that assumes the same document length, the BM25 consider not only whether a certain tag is common or rare across all documents—that is, the *inverse document frequency*—but also how frequently the tag appears in a specific document (i.e., a user or an item) associated with its "length" (Sparck Jones et al. 2000). In our case, the length of an item represents the number of unique tags annotated in the item, whereas that of a user refers to the number of unique tags used by him/her. Formally, a BM25 weight of tag t in item i, i.e., the link (t, i), is computed by:

$$W(t,i) = \log \frac{|I|}{N(i_t)} \times \frac{tf(t,i) \times (k_1 + 1)}{tf(t,i) + k_1. \times \left(1 - b + b. \frac{|i|}{avg(I)}\right)} \tag{1}$$

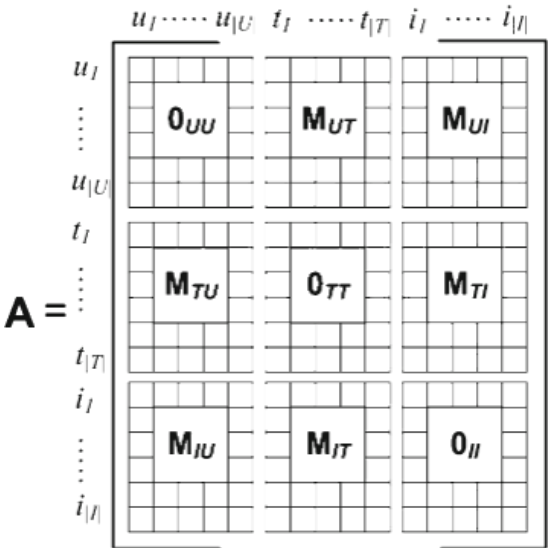

Fig. 3 A folksonomy adjacency matrix **A** in which $\mathbf{0}_{UU}$, $\mathbf{0}_{TT}$, and $\mathbf{0}_{II}$ are $|U| \times |U|$, $|T| \times |T|$, and $|I| \times |I|$ zero matrices, respectively. A $|U| \times |T|$ user-tag matrix $\mathbf{M}_{UT}$, a $|U| \times |I|$ user-item matrix $\mathbf{M}_{UI}$, and a $|T| \times |I|$ tag-item matrix $\mathbf{M}_{TI}$ are obtained from a folksonomy by aggregating over items, tags, and users, respectively

where $|I|$ is the total number of items, $N(i_t)$ is the number of items annotated with tag t, $tf(t, i)$ is the number of times item i has been tagged with tag t, $|i|$ is the length of item i, $avg(I)$ is the average length of items. Parameters k_1 and b are constants normally chosen, $k = 2.0$ and $b = 0.75$. In an analogue fashion, we can compute a BM25 weight $W(u, t)$ of tag t with respect to user u by replacing item i with user u.

4.2 Computing Katz scores

After building the adjacency matrix, we compute the Katz score for all pair of nodes in **A**. The Katz measure is one of the most effective path-based measure that has been successfully applied to different applications such as link prediction (Liben-Nowell and Kleinberg 2007), graph theory (Foster et al. 2001), and recommender systems (Huang et al. 2005). Initially the Katz score measures proximity between a pair of nodes via a weighted sum of the number of paths between the two nodes (Katz 1953). Before introducing how to calculate the Katz score between two nodes in the graph, we first define a *path* and its *path weight* in our folksonomy graph G. A path in a graph generally represents a way to get from a start node to a destination node by traversing links in the graph. Formally, A path of length n, $n \geq 1$, from a node to another node can be represented as a sequence of nodes, v_1, v_2, ..., v_n, v_{n+1} so that links are $(v_k, v_{k+1}) \in E$ for each $k = 1, ..., n$. For convenience, the weight associated to a link (v_k, v_{k+1}) is denoted using $W(v_k, v_{k+1})$. The length of a path refers to the number of links contained in the path. Therefore, if the length of a certain path is 1, there is only a direct link between two nodes. Let $P^n(x, y)$ be the set of all possible paths of length n in G from node $x$and node y. A path weight of a path $p \in P^n(x, y)$ is defined as:

$$PW_p^n(x, y) = \prod_{k=1}^{n} W(v_k, v_{k+1}) \tag{2}$$

where the initial node v_1 is node x and the terminal node v_{n+1} is node y. Then, the Katz score from x and y is calculated as:

$$K_{x,y} = \sum_{n=1}^{\infty} \alpha^n \times \left(\sum_{p \in P^n(x,y)} PW_p^n(x, y) \right) \tag{3}$$

where $\alpha \in (0, 1)$ is an attenuation parameter (usually a small value, e.g. 0.05 or 0.005). The Katz score measures the proximity of two nodes as the sum of path weights of all paths connecting the two with varying lengths. The Katz score of all pairs of nodes in the graph can be expressed in matrix form as follows (Liben-Nowell and Kleinberg 2007):

$$\mathbf{K} = \alpha\mathbf{A} + \alpha^2\mathbf{A}^2 + \alpha^3\mathbf{A}^3 + \ldots = (\mathbf{I} - \alpha\mathbf{A})^{-1} - \mathbf{I} \tag{4}$$

where $\mathbf{I}$ is the identity matrix and $\mathbf{A}$ is an adjacency matrix for the graph. As the series expansion converges if $\alpha < 1/\lambda_{\max}(\mathbf{A})$, where $\lambda_{\max}(\mathbf{A})$ is the largest absolute value of any eigenvalue of $\mathbf{A}$; thus, this condition determines how large α can be (Katz 1953). Each entry in the Katz matrix $\mathbf{K}$ represents the sum of the path weights of all lengths from the corresponding row node to the corresponding column node in the graph. We exploit the resultant matrix $\mathbf{K}$ to compute the personalized tag recommendation.

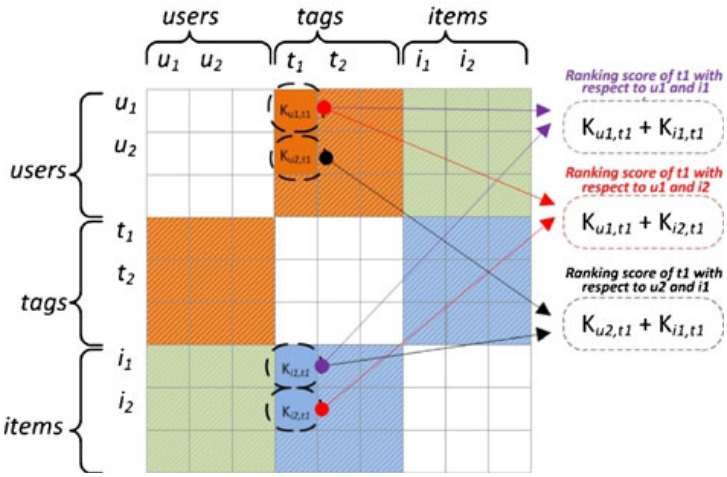

Fig. 4 The process of computing tag ranking scores based on the Katz measure

4.3 Personalized tag recommendations

In this section we present a new tag ranking method for a given user-item pair based on the **K** matrix. Since every user has a different taste on tags, we should recommend different tags to each user based on his personal interests. To identify top-N suitable tags for a given user-item pair, we assess how a certain tag is in close proximity not only to the user, but also to the item. To that end, we compute a ranking score as a sum of Katz scores. Formally, given a user u and an item i, a ranking score for a particular tag $t$is computed by:

$$score_{u,i}(t) = K_{u,t} + K_{i,t} \tag{5}$$

where $K_{u,t}$ and $K_{i,t}$ are the Katz scores from user u to tag t and from item i to tag t, respectively. We can succinctly express all ranking scores of tags for user u and item i in matrix form as:

$$\mathbf{s} = \mathbf{v}^T\mathbf{K} \tag{6}$$

where $\mathbf{v}^T$ is a vector that has value ones at entries corresponding to user u and item i, and zeros everywhere else. According to the combined Katz values contained in the resultant ranking vector **s**, the set of N ordered tags with the highest values are recommended to user u in regard to item i. Figure 4 illustrates the overall process of computing tag ranking scores of the same item i_1 for u_1 and u_2, as well as those of two different items i_1 and i_2 for the same user u_1.

5 Experiments and results

5.1 Datasets

As Table 1 summarizes, the data we used in our experiments was taken from CiteULike[1] and Last.fm[2]. CiteULike is a reference management site that can help

[1] http://citeulike.org

[2] http://www.last.fm/

Table 1 Characteristics of the datasets

Dataset	Users	Items	Tags	Tag assignments
CiteULike	2614	4,096	2310	161,395
Last.fm	1892	12,523	9749	186,479

users organize their research, collaborate with others, and discover new scientific references. We downloaded the CiteULike dataset, which is publicly available for non-commercial scientific purposes.[3] In order to focus on a dense portion of the dataset, we restrict the dataset to i) users who at least tagged 5 items and used 5 tags; ii) tags at least assigned to 5 items and used by 5 users; iii) items at least labeled by 5 tags and used by 5 users. This pruned dataset had 2614 users, 4096 items (i.e., scientific references), 2310 tags, and 161,396 tag assignments. We projected the tag assignments onto three two-dimensional matrices, resulting in 62,112 non-zero entries of the user-tag matrix (1.03 % density), 65,325 non-zero entries of the user-item matrix (0.61 % density), and 72,619 non-zero entries of the tag-item matrix (0.77 % density).

Last.fm is a social music service that assists users to discover, tag, and share music. The Last.fm dataset[4] used in this study was collected by the Informational Retrieval Groups at Autónoma University of Madrid (Cantador et al. 2011). This dataset contains 186,479 tag assignments on 12,523 items (i.e., music artists) from 1892 users with 9749 tags. We also projected the tag assignments onto three two-dimensional matrices: 35,816 non-zero entries of the user-tag matrix (0.19 % density), 71,064 non-zero entries of the user-item matrix (0.3 % density), and 109,750 non-zero entries of the tag-item matrix (0.09 % density). Note that the CiteULike evaluation dataset was relatively more clean (and dense) data compared to the Last.fm evaluation dataset. This means that the former would include few noise tags, whereas the latter would contain many personal and self-referential tags. Additionally, the CiteULike dataset did provide relatively strong connectivity between users and tags, between users and items, and between tags and items, as compared to the Last.fm dataset.

5.2 Experimental design and metrics

To evaluate the performance of our tag recommendation approach, we adapted the *leave-one-out hold-out* validation, which has widely used for evaluating tag recommender algorithms (Jäschke et al. 2008; Zanardi and Capra 2008; Rendle and Schmidt-Thieme 2010). For each user, we randomly eliminated one item and his/her tags assigned to that item from the training set. This eliminated set was used as the test set. For example, if user u_1 annotated item i_1 with two tags, t_1 and t_2, we withheld the item i_1 and the tags t_1 and t_2 from the entire tag assignments associated to user u_1. Then, by using the training set, we examined whether a tag recommender algorithm could recommend t_1 and t_2 for u_1 associated with i_1 within the top-N ranked list. We repeated this procedure five times with different test/training datasets. Thus, the reported values in the experimental results are the mean performance averaged over these five runs. In the Last.fm dataset, some users had only one tagged item

[3] http://www.citeulike.org/faq/data.adp

[4] http://www.grouplens.org/node/462

(252 out of 1892 users). We therefore did not carry out the evaluation procedure for such users, as there were no training data for them.

As evaluation metrics, we first employed the mean average precision (MAP) (Krestel et al. 2009):

$$MAP = \frac{1}{|U|}\sum_{u=1}^{|U|}\left(\frac{1}{|T_u|}\sum_{k=1}^{|T_u|} B_k \times P@k\right) \tag{7}$$

where T_u is the set of tags that user u annotated the item in the test data, $P@k$ is precision at top k, and B_k is a binary variable that is 1 if the tag with rank k in the recommended list appears in T_u and 0 otherwise.

In addition, we also measured the mean reciprocal rank (MRR) (Vallet et al. 2010) for evaluating the rank of the relevant tags:

$$MRR = \frac{1}{|U|}\sum_{u=1}^{|U|}\left(\sum_{t\in T_u}\frac{1}{rank\,(t)}\right) \tag{8}$$

where $rank(t)$ refers to the rank of relevant tag t within the ranked list for user u. If relevant tags appear at the very top of the ranked list made by an algorithm, it achieves a higher MRR value. While we focused mainly on MAP and MRR for empirical analyses, we also reported *precision* and *recall* so as to compare the performance of our method against that of other methods.

5.3 Choice of parameter $\alpha\alpha$

Before implementing our method, a crucial question is what value we should choose for the parameter α that allows for the lower efficacy of longer paths during the Katz score computation. By tuning this value, we may improve the recommendation performance. Therefore, this section investigates how our recommendation performance is sensitive to α. Because the Katz score also depends on weights on links, we assessed our performance according to the different values of α with different weights. As mentioned earlier, for convergence of the Katz score, the value of α should satisfy the condition $\alpha < 1/\lambda_{\max}(\mathbf{A})$ in which $\lambda_{\max}(\mathbf{A})$ is the largest absolute value of any eigenvalue of the adjacency matrix $\mathbf{A}$. Table 2 shows this constraint of α associated with different weights presented in Section 4.1.

We first measured MRR according to different values of α when the BM25 weight was used for computing the Katz score. As Table 2 shows, in the case of the BM25, α has to be less than 0.0069 in the CiteULike dataset and 0.0059 in the Last.fm dataset for Katz convergence; thus, we decreased the values of α from those points. Figure 5 shows MRR results on both test datasets. On the CiteULike dataset, we saw that changing α have little effect on the MRR result. On the contrary, better performance

Table 2 The largest absolute eigenvalue associated with different weights

	Weight	$\lambda_{\max}(\mathbf{A})$	$\alpha < 1/\lambda_{\max}(\mathbf{A})$
CiteULike	Binary	117.61	0.0085
	Frequency	580.70	0.0017
	BM25	144.51	0.0069
Last.fm	Binary	138.37	0.0072
	Frequency	1219.61	0.0008
	BM25	168.82	0.0059

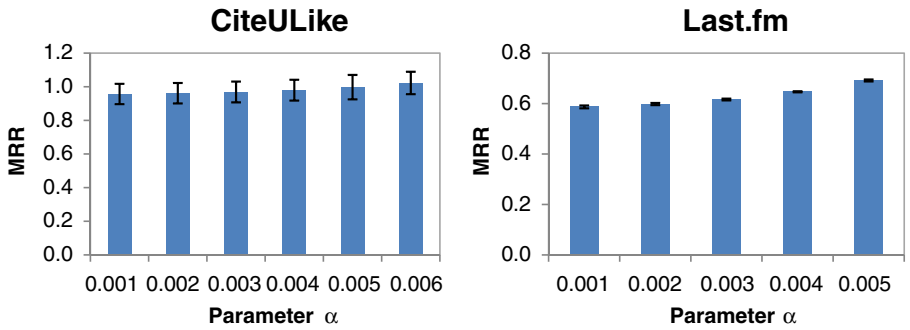

Fig. 5 MRR values according to different values of α using the BM25 weight

was achieved as α was getting closer to $1/\lambda_{\max}(\mathbf{A})$ for the Last.fm dataset. We also tested the performance when the binary and frequency weights were applied to the Katz measure. As a result, we identified that small values of α yielded slightly better performance than did higher values of α. It is worth mentioning that the choice of small α values brings long paths having less influence over the Katz score. In other words, when α becomes smaller, the Katz score is profoundly affected by shortest paths such as direct links.

We further investigated the effect of the BM25 weight in comparison with the binary and frequency weights. We chose the best value of α for each weight and measured MAP and MRR values so as to compare the results. We also tested a simply directed-link-only-based Katz score using the BM25 weight (denoted as BM25-DL) to investigate whether or not indirect connections (paths) help to improve the recommendation performance. This BM25-DL approach is analogous to a mix of most popular tags recommenders described in Jäschke et al. (2008)—that is, a mix of the most popular tags of a given user with the most popular tags of a given item—apart from the utilization of the BM25 weight. Table 3 shows the results. The second column refers to the weights used for calculating the Katz scores and the third one refers to the best value for α producing the best performance when each weight was used. The experimental results demonstrated that the Katz approach with the BM25 weight outperformed the Katz approach with the other weights on both datasets. For the CiteULike dataset, improvements of approximately 5.8 % and 3.9 % on MAP were noted compared to the binary weight and the frequency weight, respectively. However, the MRR and MAP results obtained using the BM25 and frequency weight appeared roughly the same for the Last.fm dataset. Compared to the simply BM25-DL, the Katz BM25 method that considers the ensemble of all possible paths between

Table 3 MRR and MAP values associated with different weights

	Weight	α	MAP±STDEV	MRR±STDEV
CiteULike	Binary	0.0005	0.317 ± 0.006	0.896 ± 0.058
	Frequency	0.0005	0.336 ± 0.003	0.948 ± 0.063
	BM25	0.006	0.375 ± 0.005	1.022 ± 0.066
	BM25-DL	–	0.338 ± 0.008	0.952 ± 0.058
Last.fm	Binary	0.0005	0.195 ± 0.003	0.615 ± 0.011
	Frequency	0.0001	0.254 ± 0.005	0.693 ± 0.009
	BM25	0.005	0.257 ± 0.003	0.695 ± 0.004
	BM25-DL	–	0.198 ± 0.010	0.582 ± 0.006

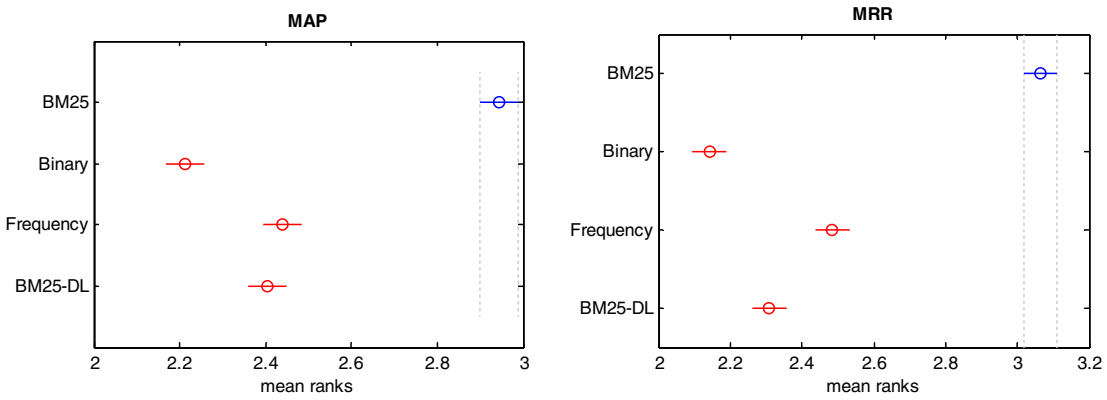

Fig. 6 Comparison of BM25 against the others with the Bonferroni test on the CiteULike dataset, shown with 95% confidence intervals

nodes yielded more precise results on both evaluation datasets. This implies that indirect paths are indeed helpful for enhancing the recommendation performance, even though longer paths have less impact on the Katz score than do shorter paths.

For statistical comparisons of multiple methods over the same test users, we conducted the Friedman test, which is a non-parametric equivalent of the repeated-measures ANOVA, as there is no guarantee for normality of recommendation accuracy distributions (Demsar 2006). If the null-hypothesis—all weighting methods perform the same and the differences found are merely random—is rejected by the Friedman test, we proceeded with the Bonferroni post-hoc test to determine which pairs of methods are significantly different, and which are not. Figures 6 and 7 depict the results of the post-hoc tests after the Friedman tests for MRR and MAP. Note that two mean ranks which are being compared are significantly different at the 5 % significance level if their confidence intervals do not overlap; otherwise, they are not significantly different. As shown, it turned out that the differences between the BM25 and frequency weights on the Last.fm dataset were not statistically significant for both MAP and MRR. Outside of this case, all the differences of the mean ranks appeared to be statistically significant at the 5 % level. Upon considering the MRR and MAP values, we chose the BM25 weigh at $\alpha = 0.006$ for the CiteULike

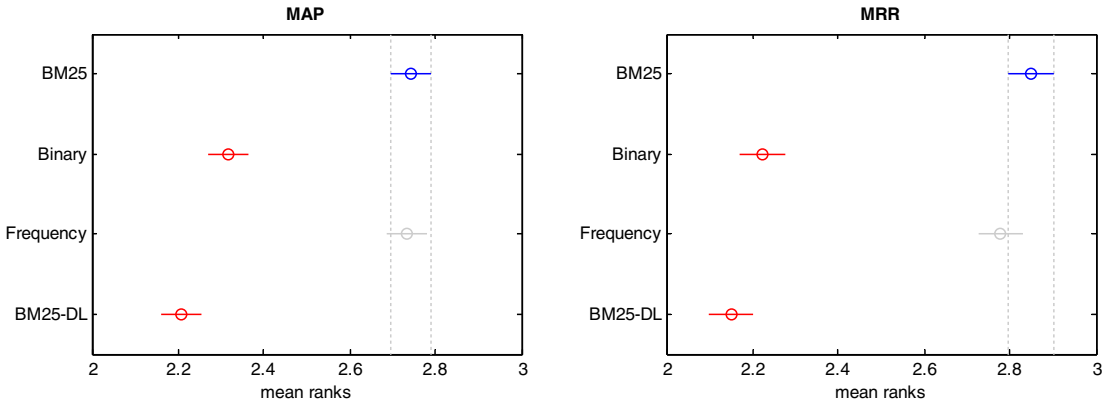

Fig. 7 Comparison of BM25 against the others with the Bonferroni test on the Last.fm dataset, shown with 95% confidence intervals

dataset and the BM25 weight at $\alpha = 0.005$ for the Last.fm dataset in the following experiments.

5.4 Comparison with other methods

In this section, we compare our method (denoted as *KaztBm25*) with the following baseline methods: (i) User-Centric Tag Model (denoted as *UCTM*) proposed by Wetzker et al. (2010), (ii) the *FolkRank* algorithm which is one of the most frequently cited studies among folksonomy-based algorithms (Jäschke et al. 2008), and (iii) the most *Popular Tag* approach (Jäschke et al. 2008). We first calculated precision and recall obtained via the four methods by changing the number of recommended tags N from 1 to 10.

Figure 8 depicts the precision-recall curves, showing how the precision and recall changes as the number of recommended tags increases. Note that the number of recommended tags is plotted on data points of the graph curves; the first point of each curve refers to the case of the top-1 whereas the last point is the case of the top-10 recommendations. Overall, we found that KatzBm25 outperformed the baseline methods in all N values on both datasets (except for recall at top-10 on the CiteULike dataset where FolkRank slightly performed better). Comparing the results of KatzBm25 and FolkRank on the CiteULike dataset, we observed that the differences became narrower as N increased. On the Last.fm dataset, however, KatzBm25 gradually outperformed the baseline methods as N increased.

We further examined MAP and MRR of each algorithm. These evaluation measures help us see whether desirable tags are appearing at the very top of the ranked list. We note that on average users had approximately 3.2 tags and 3.5 tags in the CiteULike and Last.fm test data, respectively. Consequently, on average we considered the top-3 or top-4 ranked tags when calculating MAP, because the number of recommended tags for each test user depends on how many tags he/she assigned to his/her test item. Table 4 shows MAP and MRR results obtained using the four methods. From the table, we can see that the performance of our recommender method is indeed substantially better than that of other alternatives that were considered. We also conducted Friedman tests for statistical comparisons of multiple methods in regard to MAP and MRR. As the null-hypothesis was rejected ($p < 0.05$),

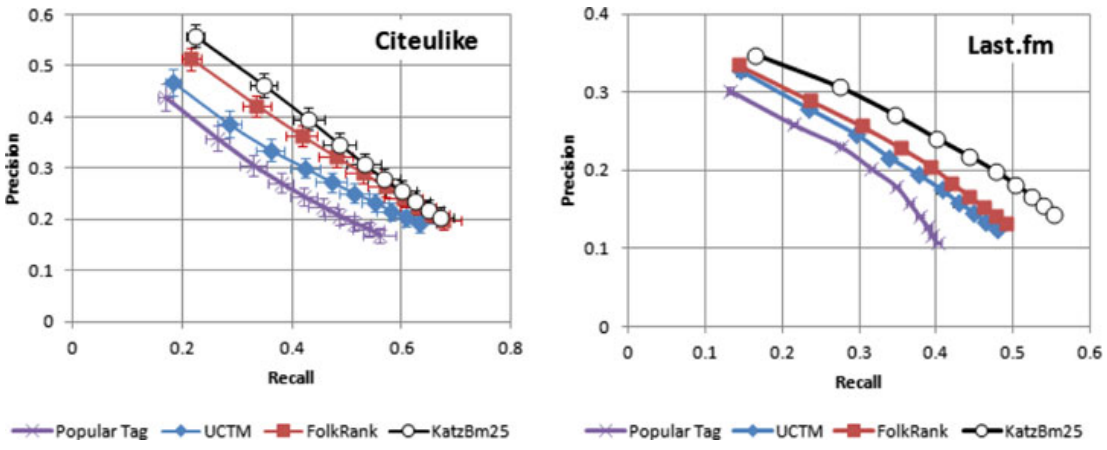

Fig. 8 Precision and recall with respect to increasing the number of recommended tags

Table 4 MRR and MAP results shown with standard deviations

	Method	MAP±STD	MRR±STD
CiteULike	Popular Tag	0.276 ± 0.008	0.808 ± 0.058
	UCTM	0.306 ± 0.007	0.902 ± 0.065
	FolkRank	0.342 ± 0.009	0.960 ± 0.062
	KatzBm25	0.375 ± 0.006	1.022 ± 0.067
Last.fm	Popular Tag	0.212 ± 0.004	0.550 ± 0.003
	UCTM	0.234 ± 0.006	0.619 ± 0.007
	FolkRank	0.226 ± 0.004	0.654 ± 0.005
	KatzBm25	0.257 ± 0.004	0.695 ± 0.009

we thus continued to perform Bonferroni post-hoc tests. As Figs. 9 and 10 show, the baseline methods have mean ranks significantly different from KatzBm25 because the confidence intervals of KatzBm25 and those of the others are all disjoint. These results confirm that MAP and MRR improvements were statistically significant.

In social tagging systems, it is often that some users are very active in utilizing tags while some other users use few tags. Accordingly, the recommendation performance for individual users could be affected by how many tags each user has used. To show this impact on the recommendation performance, for every training data, we divided users into five groups according to their tagging activities: (i) *very low taggers* who used less than 5 different tags; (ii) *low taggers* who used greater than or equal to 5 tags and less than 10 tags; (iii) *medium taggers* who used greater than or equal to 10 tags and less than 20 tags; (iv) *heavy taggers* who used greater than or equal to 20 tags and less than 40 tags; and (v) *very heavy taggers* who used greater than or equal to 40 tags. We then measured the MRR and MAP values with respect to each group of users. Table 5 summarizes each group's constraint and the average number of users who belonged to each group on both evaluation datasets.

Figures 11 and 12 show the MAP and MRR results for different groups of users on the CiteULike and Last.fm dataset, respectively. When we looked at the MAP results obtained with each method on the CiteULike dataset, every method obtained similar MAP values for each group. This was intriguing as we intuitively expected that the level of users' tagging activities could be a significant factor, influencing the quality of tag recommender systems. This result might be caused by the characteristics

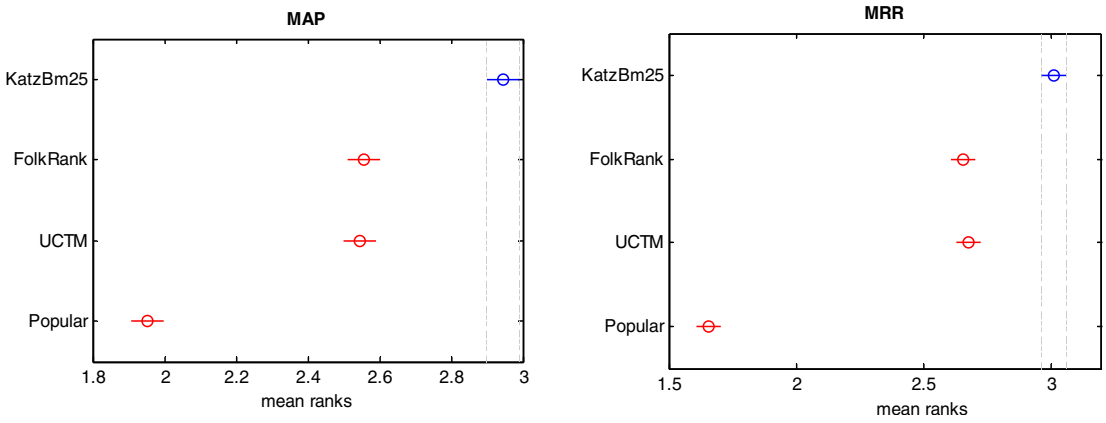

Fig. 9 Statistical comparisons of KatzBm25 against the other methods on the CiteULike dataset, shown with 95% confidence intervals

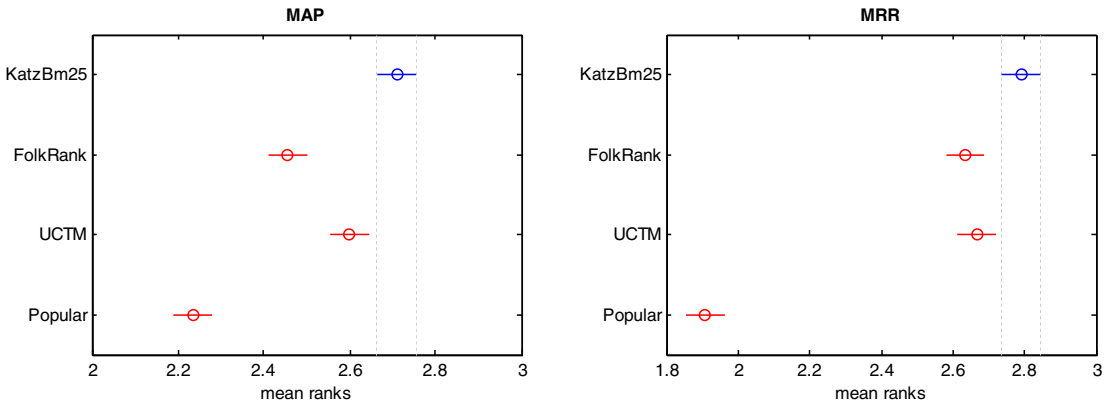

Fig. 10 Statistical comparisons of KatzBm25 against the other methods on the Last.fm dataset, shown with 95% confidence intervals

of the CiteULike dataset we experimented with. As described in Section 5.1, this dataset contained well-refined relations of a folksonomy. The performance of the tag recommender systems would not only depend on the number of tags a target user has used, but also rely on the number of tags annotated in a target item. To explain this result, we looked into this experimental data in more detail. Accordingly, we observed that on average 88.5 % of the total tags occurring in each user-item test pair also appeared in the training tag set of the corresponding user or that of the corresponding item. In other words, most of users in the refined CiteULike dataset did tend to annotate a certain item with their previously used tags or, more particularly, with tags labeled to that item. Contrary to the MAP results, the larger number of given tags per test group improved MRR for all the methods. On the Last.fm evaluation dataset, however, the performance of all the algorithms was profoundly sensitive to the number of tags used by test users as shown in Fig. 12. It was determined that the more tags users used, the better recommendation quality they received. For instance, KatzBm25 achieved an MAP of 0.33 for the very heavy taggers while achieving an MAP of 0.21 for the very low taggers (an improvement of 12 % on MAP). Similar improvements were also observed for the other methods.

We continued with comparisons to results obtained using each method within each same group. In the case of UCTM, it tended to perform well particularly for users who had many tags, but provided users having few tags with poor recommendations. On the contrary, FolkRank made good recommendations to the very lower taggers most notably on the CiteULike dataset, especially as compared to our method. As for the popular tag method, it did evidence the worst performance on all occasions,

Table 5 Distribution of the users' groups based on their tagging activities

Group of users	Users' tag usage	Average number of users	
		CiteULike	Last.fm
Very Low (VL)	num. of tags < 5	455.6	370.6
Low (L)	$5 \leq$ num. of tags < 10	595.2	305.2
Medium (M)	$10 \leq$ num. of tags < 20	556.2	332.6
Heavy (H)	$20 \leq$ num. of tags < 40	523.4	306.8
Very Heavy (VH)	$40 \leq$ num. of tags	483.6	324.8

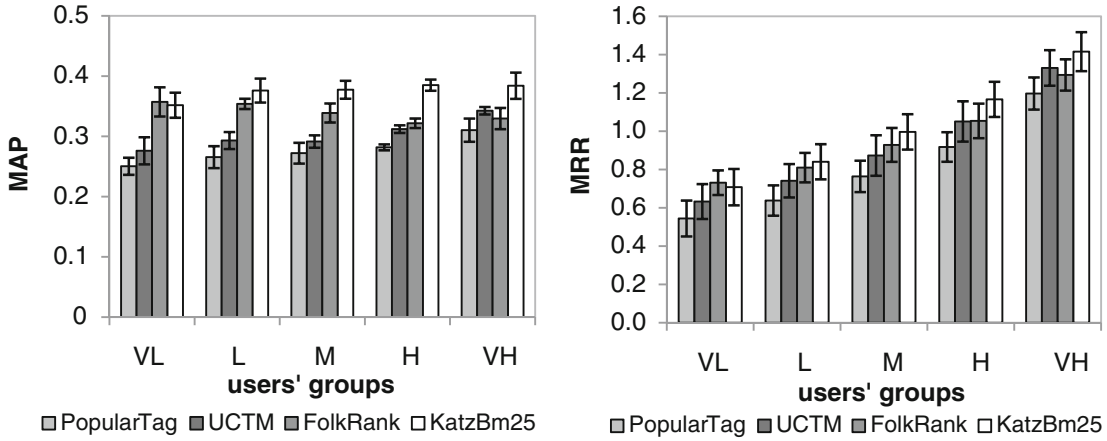

Fig. 11 The MAP and MRR result at different groups on the CiteULike dataset

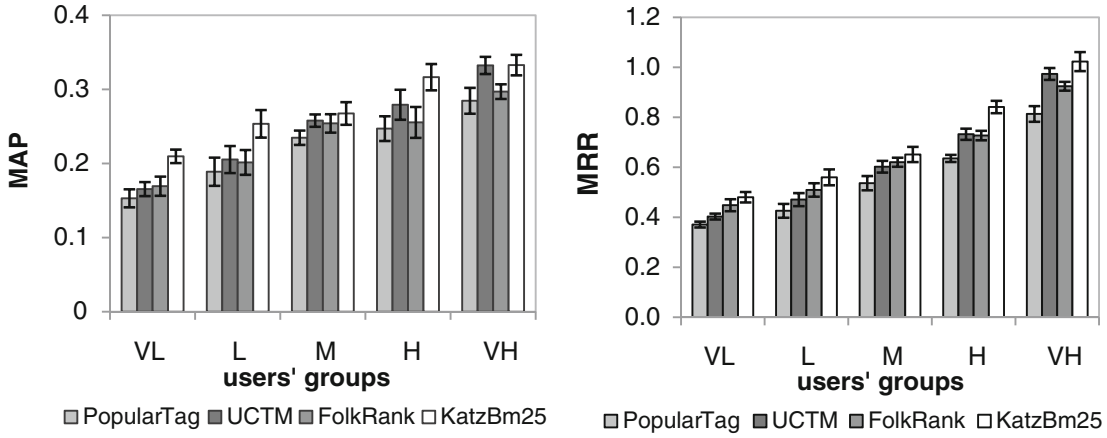

Fig. 12 The MAP and MRR result at different groups on the Last.fm dataset

Table 6 Balanced two-way ANOVA results on both datasets with respect to MAP

	Source	SS	DF	MS	F	*p*-value
CiteULike	Method	7.083	3	2.3609	66.39	.0000
	Group	0.679	4	0.1697	4.77	.0008
	Interaction	1.976	12	0.1646	4.63	.0000
	Error	141.532	3980	0.0356		
	Total	151.269	3999			
Last.fm	Method	2.723	3	0.9076	15.91	.0000
	Group	14.868	4	3.7169	65.17	.0000
	Interaction	0.457	12	0.0381	0.67	0.7837
	Error	226.995	3980	0.0570		
	Total	245.043	3999			

Table 7 Balanced two-way ANOVA results on both datasets with respect to MRR

	Source	SS	DF	MS	F	*p*-value
CiteULike	Method	29.9	3	9.9668	53.68	.0000
	Group	260.79	4	65.1972	351.16	.0000
	Interaction	3.92	12	0.3267	1.76	0.0491
	Error	738.94	3980	0.1857		
	Total	1033.55	3999			
Last.fm	Method	14.61	3	4.8695	20.37	.0000
	Group	179.71	4	44.9276	187.95	.0000
	Interaction	1.32	12	0.1103	0.46	.9374
	Error	951.39	3980	0.239		
	Total	1147.03	3999			

thereby indicating that merely the popularity of tags annotated in a given item is not enough to fully reflect individual tagging behaviors. For both two evaluation sets, KatzBm25 was consistently and clearly superior to the other methods in all cases of groups in terms of both MAP and MRR. That is, not only does our method perform well for the active users, but it is also capable of recommending more appropriate tags to the cold start users compared to the baseline methods. Since the recommendation problem for the cold start users is one of notable challenges in the field of recommender systems, the proposed approach can be beneficial to this problem.

We finally performed balanced two-way ANOVA tests for determining the effects of two independent factors at the same time (Box et al. 2005). One factor has four levels—i.e., KatzBm25, FolkRank, UCTM, and Popular Tag (recommender methods)—and another factor has five levels—i.e., VL, L, M, H, and VH (group types). To make balance ANOVA designs (i.e., the sample sizes in each group are equal), we selected 200 users per each group in the CiteULike and Last.fm datasets, respectively (1000 users in total for each dataset). Summaries of the principal ANOVA results are shown in Tables 6 and 7. From the results, we saw that both two factors, recommender methods and user groups, affected the performance of tag recommendations. In addition, there were interaction effects between the two factors in the the CiteULike dataset, whereas no interaction effects of the two were oberserved in the Last.fm dataset.

6 Conclusions and future work

In this paper, we presented a tripartite graph-based approach to identifying a list of tags that are personally tailored toward a user's interests for a given item. More precisely, the proposed approach estimates proximity between users and tags, and between items and tags based on the Katz measure, and thus discovers new triangle graphs that are likely to appear within a given folksonomy. Our experiments with the CiteULike and Last.fm datasets demonstrate that not only can the proposed method accurately recommend suitable tags for individual users, but it is also able to successfully position such tags at higher ranks. Additionally, our method is found to be fruitful in improving the recommendation quality for both the active taggers and the cold-start taggers, especially as compared to existing alternatives.

Since a folksonomy is 3-dimensional data (i.e. users, tags, and items), it is a natural direction to exploit a folksonomy tripartite structure for different applications, such as item recommendations, personalized searches, and so on. We are particularly interested in applying the Katz model on a tripartite graph space to social media annotation. Rather than recommending tags during folksonomy development, it would be interesting to uncover latent tags that have not previously been annotated in a particular item, but are likely to fit the item. Within the ambit of the tag recommendation problem, we plan to analyze directed graph link-structure to see whether the recommendation quality of our method could benefit from such a structure.

Acknowledgement This publication was made possible by a grant from the Qatar National Research Fund NPRP 09-052-5-003.

References

Bischoff, K., Firan, C.S., Nejdl, W., Paiu, R. (2008). Can all tags be used for search? In *Proceedings of the 17th ACM conference on information and knowledge management* (pp. 193–202). New York: ACM.

Box, G.E., Hunter, J.S., Hunter, W.G. (2005). *Statistics for experimenters: Design, innovation, and discovery,* 2nd ed. Wiley.

Budura, A., Michel, S., Cudré-Mauroux, P., Aberer, K. (2009). Neighborhood-based tag prediction. In *Proceedings of 6th European semantic web conference* (pp. 608–622). Berlin: Springer-Verlag.

Bu, J., Tan, S., Chen, C., Wang, C., Wu, H., Zhang, L., et al. (2010). Music recommendation by unified hypergraph: Combining social media information and music content. In *Proceedings of the international conference on multimedia* (pp. 391–400). New York: ACM.

Cantador, I., Brusilovsky, P., Kuflik, T. (2011). Second workshop on information heterogeneity and fusion in recommender systems. In *Proceedings of the fifth ACM conference on recommender systems* (pp. 387–388). New York: ACM.

Demsar, J. (2006). Statistical comparisons of classifiers over multiple data sets. *Journal of Machine Learning Research, 7,* 1–30.

Foster, K., Muth, S., Potterat, J., Rothenberg, R. (2001). A faster Katz status score algorithm. *Computational & Mathematical Organization Theory, 7*(4), 275–285.

Font, F., Serra, J., Serra, X. (2012). Folksonomy-based tag recommendation online audio clip sharing. In *Proceedings of 13th international conference on music information retrieval (ISMIR)* (pp. 73–78).

Gemmell, J., Schimoler, T., Ramezani, M., Mobasher, B. (2009). Adapting k-nearest neighbor for tag recommendation in folksonomies. In *Proceedings of the 7th workshop on intelligent techniques for web personalization & recommender systems*.

Gemmell, J., Schimoler, T., Ramezani, M., Mobasher, B., Burke, R. (2011). Tag-based resource recommendation in social annotation applications. In *Proceedings of the 6th European semantic web conference on the semantic web: research and applications* (pp. 195–206). Berlin: Springer-Verlag.

Guan, Z., Bu, J., Mei, Q., Chen, C., Wang, C. (2009). Personalized tag recommendation using graph-based ranking on multi-type interrelated objects. In *Proceedings of the 32nd international ACM SIGIR conference on research and development in information retrieval* (pp. 540–547). New York: ACM.

Hamouda, S., & Wanas, N. (2011). PUT-Tag: personalized user-centric tag recommendation for social bookmarking systems. *Social Network Analysis and Mining, 1*(4), 377–385.

Horsburgh, B., Craw, S., Massie, S., Boswell, R. (2011). Finding the hidden gems: Recommending untagged music. In *Proceedings of 22nd international joint conference on artificial intelligence - volume three* (pp. 2256–2261). AAAI Press.

Hotho, A., Jäschke, R., Schmitz, C., Stumme, G. (2006). Information retrieval in folksonomies: Search and ranking. In *Proceedings of the 3rd European semantic web conference* (pp. 411–426). Berlin: Springer-Verlag.

Huang, Z., Li, X., Chen, H. (2005). Link prediction approach to collaborative filtering. In *Proceedings of the 5th ACM/IEEE-CS joint conference on digital libraries* (pp. 141–142). New York: ACM.
Jäschke, R., Marinho, L., Hotho, A., Schmidt-Thieme, L., Stumme, G. (2008). Tag recommendations in social bookmarking systems. *AI Communications, 21*(4), 231–247.
Katz, L. (1953). A new status index derived from sociometric analysis. *Psychometrika, 18*(1), 39–43.
Krestel, R., Fankhauser, P., Nejdl, W. (2009). Latent dirichlet allocation for tag recommendation. In *Proceedings of the third ACM conference on recommender systems* (pp. 61–68). New York: ACM.
Levy, M., & Sandler, M. (2009). Music information retrieval using social tags and audio. *IEEE Transactions on Multimedia, 11*(3), 383–395.
Liben-Nowell, D., & Kleinberg, J. (2007). The link-prediction problem for social networks. *Journal of the American Society for Information Science and Technology, 58*(7), 1019–1031.
Lipczak, M., & Milios, E. (2010). Learning in efficient tag recommendation. In *Proceedings of the fourth ACM conference on recommender systems* (pp. 167–174). New York: ACM.
Liu, D., Hua, X.S., Yang, L., Wang, M., Zhang, H.J. (2009). Tag ranking. In *Proceedings of the 18th international conference on world wide web* (pp. 351–360). New York: ACM.
Miotto, R., & Orio, N. (2012). A probabilistic model to combine tags and acoustic similarity for music retrieval. *ACM Transactions on Information Systems, 30*(2), 1–29, article no. 8.
Ramezani, M. (2011). Improving graph-based approaches for personalized tag recommendation. *Journal of Emerging Technologies in Web Intelligence, 3*(2), 168–176.
Rendle, S., & Schmidt-Thieme, L. (2010). Pairwise interaction tensor factorization for personalized tag recommendation. In *Proceedings of the 3rd international conference on web search and web data mining* (pp. 81–90). New York: ACM.
Song, Y., Zhang, L., Giles, C. (2011). Automatic tag recommendation algorithms for social recommender systems. *ACM Transaction on the Web, 5*(1), 1–31.
Sparck Jones, K., Walker, S., Robertson, S.E. (2000). A probabilistic model of information retrieval: development and comparative experiments. *Information Processing and Management, 36*(6), 809–840.
Symeonidis, P., Ruxanda, M., Nanopoulos, A., Manolopoulos, Y. (2008). Ternary semantic analysis of social tags for personalized music recommendation. In *Proceedings of 9th International conference on Music Information Retrieval (ISMIR)* (pp. 219–224).
Tatli, I., & Birturk, A. (2011). A tag-based hybrid music recommendation system using semantic relations and multi-domain information. In *Proceedings of IEEE 11th international conference on data mining workshops (ICDMW)* (pp. 548–554).
Vallet, D., Cantador, I., Joemon, J. (2010). Personalizing web search with folksonomy-based user and document profiles. In *Proceedings of the 32nd European conference on advances in information* (pp. 420–431). Berlin: Springer-Verlag.
Wetzker, R., Zimmermann, C., Bauckhage, C., Albayrak, S. (2010). I tag, you tag: Translating tags for advanced user models. In *Proceedings of the 3rd ACM international conference on web search and data mining* (pp. 71–80). New York: ACM.
Xu, S., Bao, S., Fei, B., Su, Z., Yu, Y. (2008). Exploring folksonomy for personalized search. In *Proceedings of 31st annual international ACM SIGIR conference* (pp. 155–162). New York: ACM.
Zanardi, V., & Capra, L. (2008). Social ranking: Uncovering relevant content using tag-based recommender systems. In *Proceedings of 2nd ACM conference on recommender systems* (pp. 51–58). New York: ACM.