Multiple Interests of Users in Collaborative Tagging Systems

Ching-man Au Yeung, Nicholas Gibbins and Nigel Shadbolt

Abstract Performance of recommender systems depends on whether the user profiles contain accurate information about the interests of the users, and this in turn relies on whether enough information about their interests can be collected. Collaborative tagging systems allow users to use their own words to describe their favourite resources, resulting in some user-generated categorisation schemes commonly known as folksonomies. Folksonomies thus contain rich information about the interests of the users, which can be used to support various recommender systems. Our analysis of the folksonomy in Delicious reveals that the interests of a single user can be very diverse. Traditional methods for representing interests of users are usually not able to reflect such diversity. We propose a method to construct user profiles of multiple interests from folksonomies based on a network clustering technique. Our evaluation shows that the proposed method is able to generate user profiles which reflect the diversity of user interests and can be used as a basis of providing more focused recommendation to the users.

1 Introduction

As the volume of information available on the Web continues to grow at a dramatic rate, recommender systems [1] are becoming increasingly desirable. While Web users find it difficult to locate information relevant to their needs, information providers also find it difficult to deliver their information to the target audience. A recommender system can be used to solve the problem by filtering information on behalf of the users and recommending potentially interesting resources to them.

A crucial element in a recommender system is the representation of user interests, which is usually referred to as a user profile [7]. The performance of a recommender system greatly depends on whether the user profiles truly reflect user interests.

C.M. Au Yeung (✉)
Intelligence, Agents, Multimedia Group, School of Electronics and Computer Science, University of Southampton, Southampton, SO17 1BJ, UK
e-mail: cmay06r@ecs.soton.ac.uk

I. King, R. Baeza-Yates (eds.), *Weaving Services and People on the World Wide Web*,
DOI 10.1007/978-3-642-00570-1_13,

While some research works attempt to construct user profiles based on the browsing history of the users [9, 27], some generate user profiles by analysing the documents collected by the users [4].

In recent years, the rising popularity of collaborative tagging systems such as Delicious provide new sources of information about the interests of Web users. Collaborative tagging systems [8] allow users to choose their own words (tags) to describe their favourite Web resources, resulting in an emerging classification scheme now commonly known as a ***folksonomy*** [28]. Given that the resources and the tags posted by Web users to these systems are highly dependent on their interests, folksonomies thus provide rich information for building more accurate and more specific user profiles for use in recommender systems.

There have been only a few studies in the literature which try to model user interests based on the information available in collaborative tagging systems [5, 17], and usually only a single set of frequently used tags are obtained to represent user interests. However, we observe that tags used by users are very diverse, implying that users have a wide range of interests. Hence, a single set of tags is not the most suitable representation of a user profile because it is not able to reflect the multiple interests of users. In this chapter, we propose a network analysis technique performed on the personomy [12] of a user to identify the different interests of the user, and to construct a more comprehensive user profile based on the results.

The remaining of this chapter is structured as follows. In Sect. 2 we briefly introduce folksonomies and personomies. Section 3 presents our analysis of the personomies obtained from Delicious. We describe our proposed method for generating user profiles from personomies in Sect. 4. Section 5 presents results of our evaluation and discusses the usefulness of the generated user profiles. We mention related works in Sect. 6 and finally give concluding remarks and future research directions in Sect. 7.

2 Folksonomies and Personomies

In a collaborative tagging system, users are allowed to choose any terms they like to describe their favourite Web resources. Folksonomies [28] represent user-contributed metadata aggregated in these systems. A folksonomy is generally considered to consist of at least three sets of elements, namely users, tags and documents [11, 16, 19].

Definition 1. A folksonomy $\mathbf{F}$ is a tuple $\mathbf{F} = (U, T, D, A)$, where U is a set of users, T is a set of tags, D is a set of Web documents, and $A \subseteq U \times T \times D$ is a set of annotations.

A folksonomy can be sliced into different sub-parts depending on which kind of elements one focuses on. In this chapter, we focus on the users and are interested in the collections of tags and documents possessed by individual users, which are

given the name ***personomy*** [12].[1] In order to extract the set of tags and documents associated with a user, we can slice a folksonomy by narrowing our attention to a particular user.

Definition 2. A personomy $\mathbf{P}_u$ of a user u is a restriction of a folksonomy $\mathbf{F}$ to u: i.e. $\mathbf{P}_u = (T_u, D_u, A_u)$, where A_u is the set of annotations of the user: $A_u = \{(t, d)|(u, t, d) \in A\}$, T_u is the user's set of tags: $T_u = \{t|(t, d) \in A_u\}$, and D_u is the user's set of documents: $D_u = \{d|(t, d) \in A_u\}$.

A personomy can be represented in the form of a graph with nodes representing the tags and documents associated with this particular user. If folksonomy can be considered as a hypergraph with three disjoint sets of nodes (user, tags and documents), a personomy can be represented as a bipartite graph with two disjoint sets of nodes. The bipartite graph TD_u of a personomy of a user u is defined as follows.

$$TD_u = \langle T_u \cup D_u, E_{td} \rangle, E_{td} = \{(t, d)|(t, d) \in A_u\}$$

In other words, an edge exists between a tag and a document if the tag is assigned to the document by the user. The graph can be represented in matrix form, which we denote as $\mathbf{X} = \{x_{ij}\}$, $x_{ij} = 1$ if there is an edge connecting t_i and d_j, and $x_{ij} = 0$ otherwise.

We can further fold the bipartite graph into a one-mode network [19] of documents: $\mathbf{A} = \mathbf{X}^{\mathrm{T}}\mathbf{X}$. The adjacency matrix $\mathbf{A} = \{a_{ij}\}$ represents the personal repository of the user. a_{ij} represents the number of tags which have been assigned to both documents d_i and d_j. Thus, documents with higher weights on the edges between them can be considered as more closely related. On the other hand, a one-mode network of tags can be constructed in a similar fashion: $\mathbf{B} = \mathbf{X}\mathbf{X}^{\mathrm{T}}$. $\mathbf{B}$ represents a semantic network which consists of the associations between different tags. Tags in this network are connected by edges whose weights reflect how frequently the tags co-occur with each other. This can be considered as a simple ontology used by the particular user.

3 Analysis of Personomies

In order to understand the characteristics of personomies in collaborative tagging systems, we carry out some analyses on the personomies collected from Delicious.[2] Delicious is a social bookmarking site which allows users to assign tags to bookmarks. We use a crawler program written in Python to collect data of Delicious users in the period between December 2007 and February 2008. As there are some users who have assigned no tags to any of their bookmarks on the system, these users are

[1] In the blogosphere, the term personomy has also been used in a more general sense to represent the aggregated digit manifestation of a user on the Web. See http://personomies.com/what-are-personomies/.

[2] http://delicious.com

filtered out when performing the following analysis. The dataset after filtering contains a total of 9,185 unique users, with 514,929 unique tags and 3,281,306 unique bookmarks.

3.1 Number of Tags and Documents of a User

Firstly, we take a look at the number of tags and documents possessed by the users. On average a user has 285 unique tags and has 602 unique bookmarks on Delicious. Although some users have over 18,000 tags and over 34,000 bookmarks, only a very small number of users have more than 1,000 tags or bookmarks. Figure 1 shows graphs of the number of tags and bookmarks of the users. The graphs in logarithmic scale show that the distribution of frequencies of tags and bookmarks follows the power law. This finding agrees with what Golder and Huberman [8] report in one of the earliest papers on collaborative tagging systems, showing that there are a small number of users having a large number of tags and bookmarks, and a large number of users having a small number of tags and bookmarks.

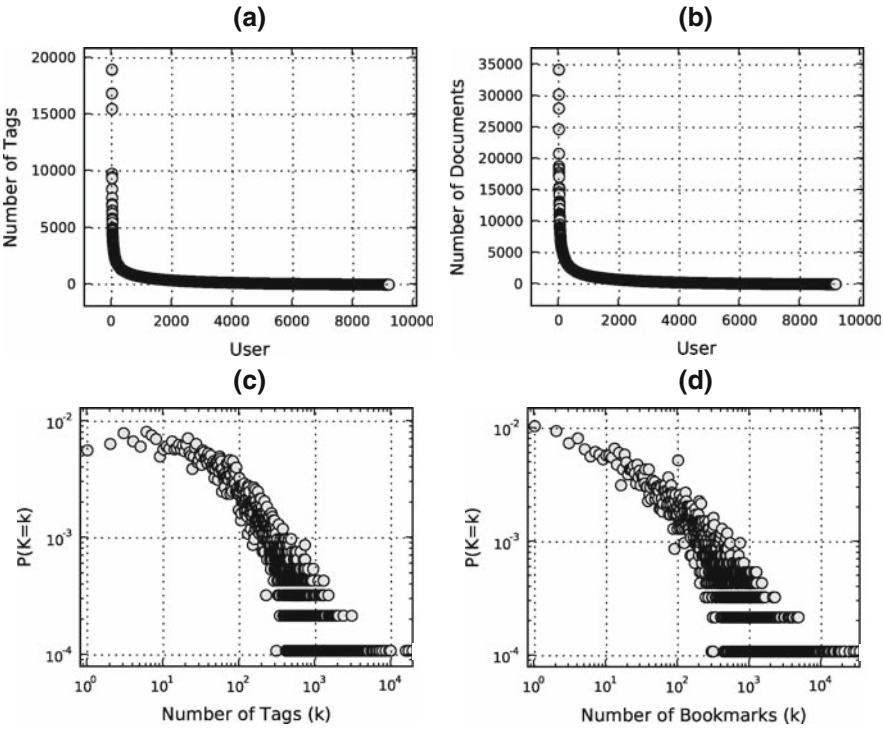


Fig. 1 Graphs showing the number of tags and bookmarks of the users in Delicious. (**a**) and (**b**) plots the data by sorting the users according to the number of tags and bookmarks they have. (**c**) and (**d**) are plots in logarithmic scale showing the distribution of the frequencies

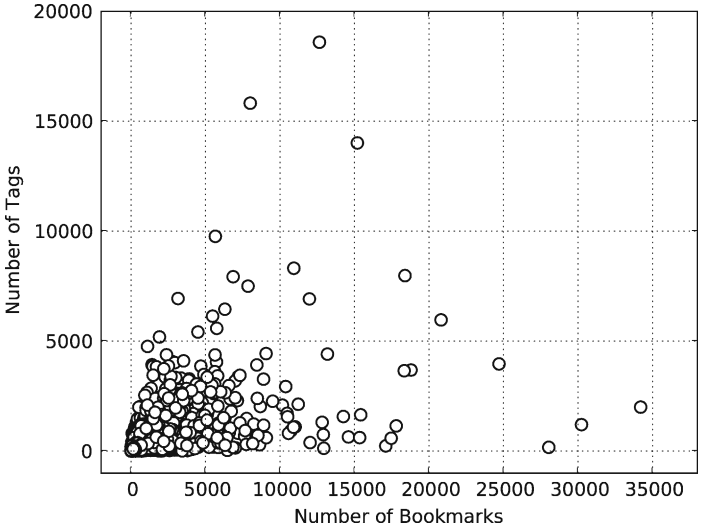


Fig. 2 Scatter plot of the number of distinct tags against the number of distinct bookmarks of the users

Secondly, we examine the correlation between the number of tags and the number of bookmarks of the users. Figure 2 shows a scatter plot of the data. It reveals a moderate relationship between the number of tags and the number of bookmarks, with a correlation coefficient of 0.55. Tagging can be considered as a kind of indexing. When there are more bookmarks in the collection of a user, the number of bookmarks appearing under any particular tag will also tend to increase. Hence it is natural to assume that the number of unique tags used by the user will also increase because it becomes necessary to use more tags to distinguish between different bookmarks by putting them into more specific categories.

However, tagging is also a very personal and subjective way of categorising bookmarks. The bookmarks and tags of the users are actually highly dependent on the interests of the users. If a user has a very specific interest, a small number of tags will be enough for even a large collection of bookmarks, because they will probably be about the same topic. On the other hand, if a user has diverse interests, more tags may be required to describe even a small number of bookmarks.

A further investigation of the data reveals that the correlation between the two numbers appears to be stronger for users with fewer bookmarks than those with many bookmarks. For users with fewer than 500 bookmarks, a correlation coefficient of 0.43 is obtained. For users with more than 5,000 bookmarks, the correlation coefficient is only 0.14. A similar result can also be found in [8]. This may suggest that users with many bookmarks can behave very differently: while some may stick to using a small number of tags on new bookmarks, others may continue to introduce new tags.

3.2 Measuring Diversity of Interests

The diversity of user interests is an important issue to be understood before we can accurately model user interests to provide recommendations. When users have multiple and diverse interests, the user profile should be able to reflect this diversity so that a recommender system will be able to provide recommendations which satisfy the different needs of the users.

Here, we propose two measures which are designed to reflect the diversity of interests of the users. We will give examples based on the two fictional users listed in Table 1, one with rather specific interests in Semantic Web related topics, while another has more diverse interests such as cooking and sports.

Our first measure involves examining how frequently a tag is used on the collection of resources of the users. Intuitively, if a user is only interested in only one or two topics, we would expect the tags used by this user to appear on most of the resources. On the other hand, if the interests of the user are very diverse, the tags are more likely to be used on only a small portion of the resources. This is because different tags are required to describe resources related to different interests of the user. To quantify this characteristic, we propose a measure called ***tag utilisation*** which is defined as follows.

Definition 3. Tag utilisation of a user u is the average of the fractions of bookmarks on which a tag is used:

$$TagUtil(u) = \frac{1}{|T_u|} \sum_{t \in T_u} \frac{|D_{u,t}|}{|D_u|} \tag{1}$$

where $D_{u,t}$ is the set of documents assigned the tag t: $D_{u,t} = \{d|(t, d) \in A_u\}$.

In addition, the diversity of a user's interest can also be understood by examining tag co-occurrence. If for a user the tags are always used together with each other, it is likely that the tags are about similar topics, and therefore it can be suggested that the user has a rather specific interest. If on the other hand the tags are mostly used separately, they are more likely to be about different topics, and thus reflect that the user has multiple interests which are quite distinctive from each other. Such characteristic can be measure by the ***average tag co-occurrence ratio***, which is defined as follows.

Table 1 Two example users with their personomies

User	Resource	Tags
u_1	r_1	web2.0, semanticweb, ontology, notes
	r_2	semanticweb, ontology
	r_3	semanticweb, ontology, rdf
u_2	r_4	semanticweb, folksonomy, tagging
	r_5	toread, cooking, recipe, food
	r_6	sports, football, news

Definition 4. Average tag co-occurrence ratio of a user measures how likely two tags are used together on the same bookmark by a user:

$$Avg\_Tag\_Co(u) = \sum_{t_i, t_j \in T_u, t_i \neq t_j} \frac{Co(t_i, t_j)}{2 \times C_2^{|T_u|}} \tag{2}$$

If we represent the co-occurrences between the tags as a network (by constructing the adjacency matrix **B**), we can see that the average tag co-occurrence ratio is actually equivalent to the density of the network of tags: $Co(t_i, t_j)$ counts the number of edges in the network, while $C_2^{|T_u|}$ calculates the number of possible edges based on the number of nodes. This agrees with the formula of the density of a network:

$$Density = \frac{2 \times |E|}{|V| \times (|V| - 1)} \tag{3}$$

where E is the set of edges and V is the set of nodes. Hence, the average tag co-occurrence ratio actually reflects the cohesion [29] of the network of tags, which in turn reflects whether the tags are related to a specific domain or a wide range of topics.

As an illustrating example, we apply these two measures to the two users listed in Table 1. The tag utilisation of u_1 is 0.60, while that of u_2 is 0.33. The average tag co-occurrence ratio of u_1 is 0.80, while that of u_2 is 0.27. For both measures, u_1 scores higher than u_2, this agrees with the fact that the interests of u_2 are more diverse as observed from this user's collection of resources.

We apply the calculations of tag utilisation and average tag co-occurrence ratio to the data collected from Delicious. The average values of these two measures of the users are plotted in Fig. 3.

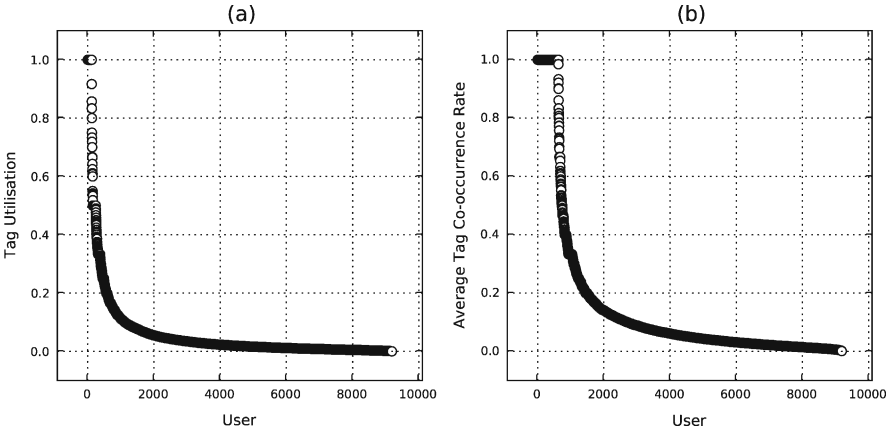


Fig. 3 Tag utilisation and average tag co-occurrence ratio of the personomies collected from Delicious. The x-axis represents the ranks of users sorted by their scores in descending order

Although the two measures consider different characteristics of the personomy of a user, the results are very similar. Firstly, there is a strong correlation between tag utilisation and average tag co-occurrence ratio, with a correlation coefficient of 0.71. The mean values of tag utilisation and average tag co-occurrence ratio are both very low, at 0.06 and 0.07 respectively, even though the values span across the whole range of 0–1 inclusively. This means that on average a tag is only used on 6% of the bookmarks in a user's collection, and that a tag is only used together with 7% of other tags. We can see that there is a small group of points in both graphs in Fig. 3(a) and (b) which attain a value of 1. These actually correspond to users who have only one bookmark in their collection. Other than these the values drop quickly, and the majority (93% in both cases) of personomies have values less than 0.2.

These figures suggest that for most users many tags are used only on a small portion of their bookmarks, and that these tags are not always used together. This shows that bookmarks collected by the users have topics which are very diverse such that a particular tag is only useful on a small portion of them and that the users keep tags which represent concepts in very different domains. Hence, this indicates that most users of Delicious have diverse interests instead of a single interest in a very specific domain.

4 Generating User Profiles of Multiple Interests

As the majority of users are observed to be interested in a wide range of topics from different domains, a user profile in the form of a single set of tags is definitely inadequate. Hence, for applications which provide different services based on user interests, it is very much desirable to have user profiles which can accommodate the multiple interests and present a more fine-grained representation of the users.

Identifying the different interests can be a challenging task as tags are freely chosen by users and their actual meaning is usually not clear. A solution to this problem is to exploit the associations between tags and documents in a folksonomy. As it is obvious that documents related to the same interest of a user would be assigned similar tags, clustering algorithms can be applied to group documents of similar topics together. We can then extract the sets of tags assigned to these documents and use them to represent the multiple interests of the users.

Based on this idea, we propose a method for constructing user profiles which involves constructing a network of documents out of a personomy, applying community-discovery algorithms to divide the nodes into clusters, and extracting sets of tags which act as signatures of the clusters to represent the interests of the users.

4.1 Community Discovery Algorithms

Clusters in a network are groups of nodes in which nodes have more connections among each other than with nodes in other clusters. The task of discovering clusters

of nodes in a network is usually referred to as the problem of discovering community structures within networks [6]. Approaches to this problem generally fall into one of the two categories, namely agglomerative, which start from isolated nodes and group nodes which are similar or close to each other, and divisive, which operate by continuously dividing the network into smaller clusters [24].

To measure the "goodness" of the clusters discovered in a quantitative way, the measure of ***modularity*** [21] is usually used. The modularity of a particular division of a network is calculated based on the differences between the actual number of edges within a community in the division and the expected number of such edges if they were placed at random. Hence, discovering the underlying community structure in a network becomes a process of optimising the value of modularity over all possible divisions of the network.

Although modularity provides a quantitative method to determine how good a certain division of a network is, brute force search of the optimal value of modularity is not always possible due to the complexity of the networks and the large number of possible divisions. Several heuristics have been proposed for optimizing modularity, these include simulated annealing [10], and removing edges based on edge betweenness [21]. In addition, a faster agglomerative greedy algorithm for optimizing modularity, in which edges which contribute the most to the overall modularity are added one after another, has been proposed [20]. In this chapter, we will employ this fast greedy algorithm to perform clustering, as it is efficient and performs well on large networks.

4.2 User Profile Generation

Given a network of documents (which are bookmarks in our case), we apply the community-discovery algorithms to obtain clusters of documents. As the different clusters should contain documents which are related to similar topics, a cluster can be considered to correspond to one of the many interests of the user. A common way to represent user interests is to construct a set of tags or a tag vector. Similarly, we can obtain a set of most frequently used tags from each of the document clusters to represent the corresponding interest. As a summary of our method, the following list describes the whole process of constructing a user profile for user u.

1. Extract the personomy $\mathbf{P}_u$ of user u from the folksonomy $\mathbf{F}$, and construct the bipartite graph TD_u.
2. Construct a one-mode network of documents (by generating the adjacency matrix $\mathbf{A}$) out of TD_u, and perform modularity optimization over the network of documents using the fast greedy algorithm.
3. For each of the clusters (communities) c_i obtained in the final division of the network, obtained a set K_i of tags which appear on more than $f\%$ of the documents in the cluster. The set of tags of a cluster is treated as a signature of that cluster.
4. Finally, return a user profile P_u in the form of a set of $K_i's$: $P_u = \{K_i\}$.

Table 2 A resultant user profile

User A	
K_1	webdesign, web2.0, tutorial, blog, css
K_2	linux, opensource, ubuntu, software
K_3	webhosting, filesharing
K_4	grammar, english
K_5	digg, sharing, music, mp3

For the signatures of the clusters, one can include all the tags which are used on the bookmarks in the cluster, or include only the tags which are common to the bookmarks in the cluster. However, the set of tags chosen for a cluster will affect how accurate the profile is in modelling the user's interest. In general, for a large value of f only the most common tags in the cluster will be included in the signature, while a small value of f will include more tags in the signature. We will investigate the problem of choosing a right value for f in the following section. As an illustrating example, Table 2 shows the result of applying the proposed method on one personomy in our data set, with $f = 20\%$ (see also Fig. 4 for the visualisation of the network of documents).

5 Evaluation and Discussions

We believe that the use of multiple sets of tags in user profiles should give a more accurate representation of the interests of the users. It is also our hypothesis that user profiles generated by the proposed method will reveal the multiple interests of a user such that recommender systems using these profiles will be able to serve the different interests of the user better. Therefore we try to evaluate our proposed method by asking the following two questions. Firstly, are the sets of tags extracted from the clusters accurate descriptions of the bookmarks from which they are extracted?

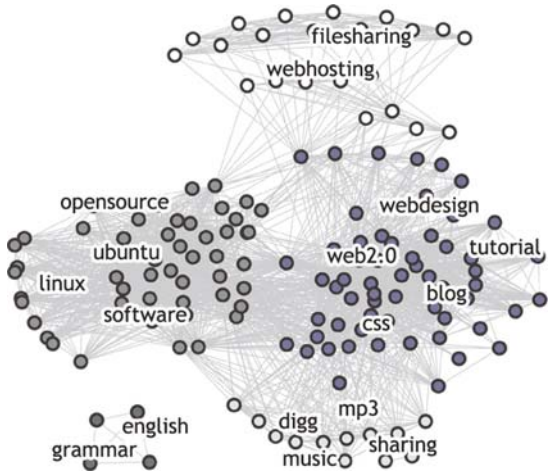


Fig. 4 An example of clustering of a personomy

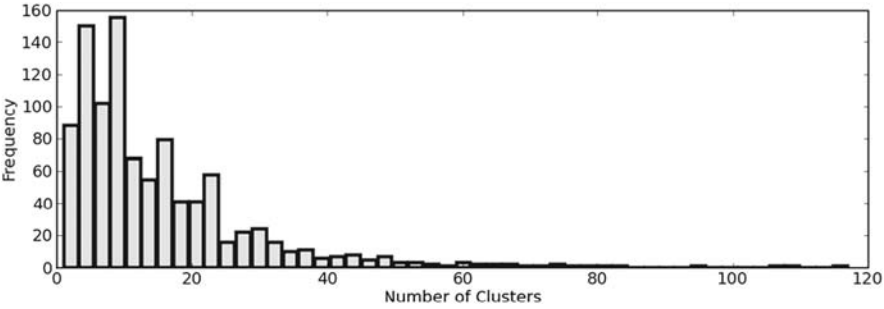


Fig. 5 Number of clusters discovered for the 1,000 personomies

If this is the case, then the user profiles should accurately represent the interests of the users. Secondly, can the generated user profiles be used to retrieve more relevant items than other user profiles which are generated without considering the multiple interests of the users? The two parts of our evaluation attempt to answer these two questions respectively.

To start with, we select at random 1,000 users from our dataset who have over 100 bookmarks in their personomies. The requirement of having at least 100 bookmarks is to ensure that there are enough bookmarks for clustering so that clearer results can be obtained. We apply our proposed method of generating user profiles on these personomies, and obtain a set of clusters of bookmarks and their signatures. We discover that there are a substantial number of clusters with only one bookmark. The bookmarks in these clusters are mostly not assigned any tags. Hence, we exclude these single-bookmark clusters in the following analysis. Figure 5 graphs the number of clusters discovered for each of the personomies. On average 15 clusters are discovered for a personomy in the dataset.

5.1 Representation of User Profiles

Our first question concerns with the issue of whether the sets of tags in the user profile are accurate descriptions of the bookmarks in the clusters. An appropriate method of evaluation is to approach this question from an information retrieval perspective. Given the signature of a cluster as a query, can we retrieve all the bookmarks within that cluster and avoid obtaining bookmarks in other clusters which are irrelevant? In addition, how many tags should be included in the signature in order to accurately describe a cluster? To answer such questions, we employ the measures of precision and recall [25] which are commonly used for evaluating information retrieval systems.

Precision and recall are two widely used measures for evaluating performance of information retrieval. Precision measures the fraction of documents in the retrieved set which are relevant to the query, while recall measures the fraction of relevant documents that the system is able to retrieve.

To employ the precision and recall measures, we treat the signatures of the clusters as queries, and use them to retrieve bookmarks by comparing the tags assigned to them to those in the queries. As for the representation of tags, we employ a vector space model of information retrieval. In other words, for each personomy, we construct a term vector $e = (e_1, e_2, \ldots, e_n)$ for each bookmark, with $e_i = 1$ if the bookmark is assigned the ith tag, and $e_i = 0$ otherwise. Similarly, the signature of a cluster is converted into a query in the form of a term vector q. The retrieval process is carried out by calculating the cosine similarity between the query vector and the bookmark vectors:

$$Sim(q, e) = \frac{q \cdot e}{|q||e|} \tag{4}$$

Those with similarity higher than a certain threshold t will be retrieved ($0 \leq t \leq 1$). For a cluster c, let the set of bookmarks in the cluster be D_c, and the set of bookmarks retrieved by the signature of the cluster be D_x. The precision and recall of the system on c are defined as follows. In addition, we also consider the F_1 measure [25] which is a combined measure of precision and recall.

$$Precision(c) = \frac{|D_x \cap D_c|}{|D_x|} \tag{5}$$

$$Recall(c) = \frac{|D_x \cap D_c|}{|D_c|} \tag{6}$$

$$F_1(c) = \frac{2 \times Precision(c) \times Recall(c)}{Precision(c) + Recall(c)} \tag{7}$$

We calculate the three measures for the user profiles generated from the 1,000 selected personomies. The results are presented in Fig. 6. We control two parameters in our evaluation. The first parameter is f (tag threshold in Fig. 6), the percentage of bookmarks above which a tag is assigned to in a cluster for it to be included in the signature. The second one is t, the threshold of cosine similarity.

Figure 6(a) shows that for most values of similarity threshold precision attains maximum for f in the range from 0.1 to 0.4, and thereafter it continues to decrease as f increases. The result suggests that if only the most common tags are included in the signatures, they will become less representative as summaries of the clusters. This is probably due to the fact that the most common tags are usually too general and a query constructed from these tags will tend to retrieve bookmarks from other clusters as well which are related to a different sub-topic under the common tags. On the other hand, when one includes all the tags which appear in a cluster (with $f = 0\%$), the signature will include too many tags such that it will not be similar to any of the signatures of the bookmarks, leading again to a low precision.

As for recall, we observe some differences for different values of similarity threshold. For small values of t (from 0.0 to 0.3), recall continues to decrease as f increases. However, for larger values of t (from 0.4 to 1.0), recall first increases and then decreases as t increases. This is probably due to the reason that when

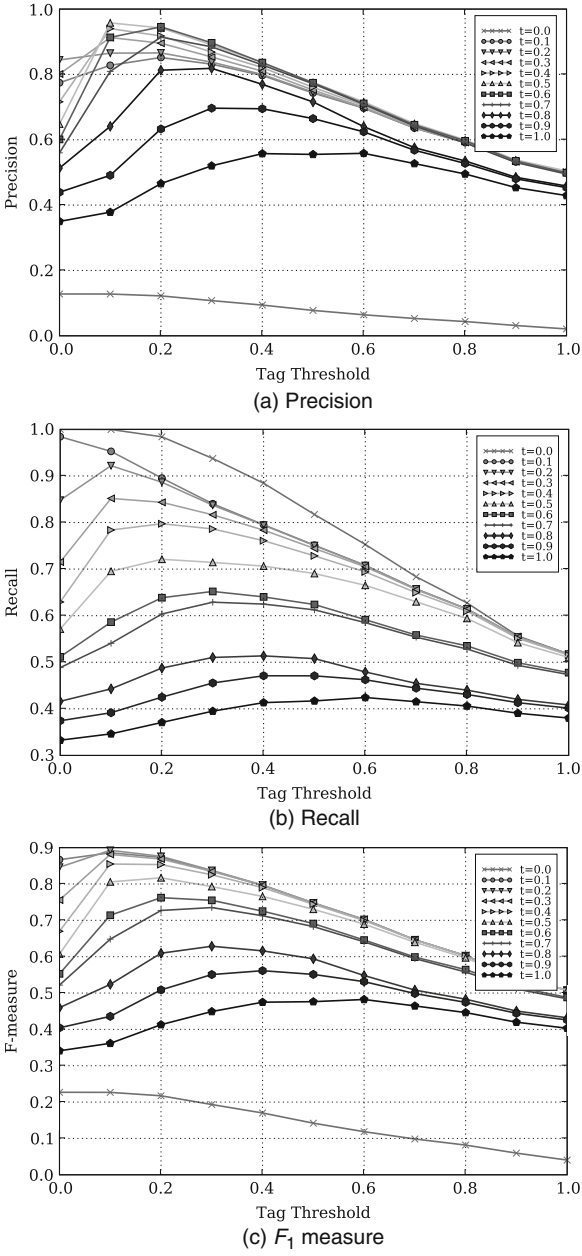


Fig. 6 Precision, recall and F_1 measure. Different lines correspond to different values of similarity threshold

the similarity threshold is low, the number of tags in the cluster signature is less important as most of the bookmarks will be retrieved even if their similarity with the query is small. As f increases, fewer tags are included in the signature and therefore it becomes more difficult to retrieve relevant bookmarks. On the other hand, when

t becomes higher, signatures which include all the tags in a cluster or include only the most common tags are very dissimilar to any of the bookmarks in the cluster, therefore recall attains maximum somewhere between the two extremes.

For common values of similarity threshold between $t = 0.3$ and $t = 0.5$, precision and recall attain maximum for values of f between 0.1 and 0.2, with precision over 0.8 and recalls over 0.7. F_1 measures also attain maximum around these values of t and f. This suggests that it is better to include more tags in a cluster signature so as to make it specific enough for representing the topic of the cluster (and thus the interest of the user represented by the cluster). Given these results, we conclude that by choosing a suitable value of f the tags extracted do constitute good descriptions of the bookmarks within the clusters.

5.2 Usefulness of the Generated User Profiles

Our second question concerns whether the user profiles generated by our proposed method will provide better support to recommender systems. To answer this question, we divide our data into a training set and a test set. We extract the first 70% bookmarks and the tags associated with them, and use them to generate a profile for the user using our proposed method. The generated user profile is then used to retrieve the remaining 30% of the bookmarks in the user's personomy. The bookmarks are retrieved according to the similarity between the sets of tags in the user profile and the tags assigned to the bookmarks. In these experiments we employ the following similarity measure between two sets of tags. This measure is chosen because it gives more distinguishable values when the similarity is low, which is common when a bookmark is assigned a large number of tags.

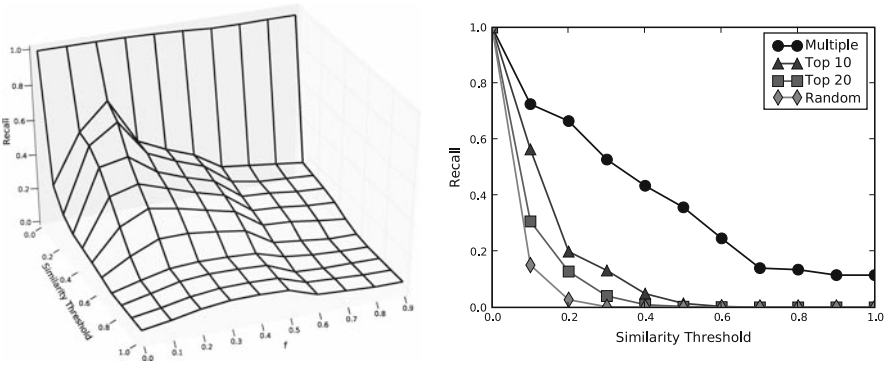
$$Sim(X, Y) = \frac{2 \times |X \cap Y|}{|X| + |Y|} \tag{8}$$

We again adopt the notion of recall as a performance measure to judge the usefulness of the generated user profiles. Let D_i^α be the set of bookmarks retrieved by the user profiles at the similarity threshold α ($0 \leq \alpha \leq 1$), and D_r be the set of bookmarks in the test set, recall is then defined as follows.

$$Recall(\alpha) = \frac{|D_i^\alpha \cap D_r|}{|D_r^\alpha|} \tag{9}$$

Our evaluation involves two experiments. In the first one is aimed at determining the optimal value of f, the fraction of tags to be included in the signature of a cluster, at which the user profiles are best at retrieving or recommending bookmarks which are interesting to the users. The result of this experiment is shown in Fig. 7(a).

Figure 7(a) plots recall against different values of similarity threshold for different values of f. The result shows that the user profiles do not help retrieve relevant bookmarks when too few tags, i.e. large values of f, are included in the signatures of



(a) (b)

Fig. 7 (**a**) Recall at different values of f. (**b**) Comparing the level of recall when different types of user profiles are used

the clusters. The optimal value of $f \simeq 0.2$ means that more tags should be included in the signature for better recall. This suggests that while each cluster might be characterised by one or two tags which represent the main topic of that cluster, there are also other tags which helps to describe bookmarks belonging to a sub-category. As a result, a signature should also include these tags such that it is able to retrieve more relevant bookmarks. This actually agrees with the results we presented in the previous section.

Figure 7(a) also shows that the graph flattens as the similarity threshold goes beyond 0.7. When we take a closer look at the results, we find that at these points most of the signatures only consist of one tag, meaning that they are only able to retrieve bookmarks which have been assigned that tag. Hence, recall experiences very few changes beyond that point.

In the second experiment, we compare the user profiles generated by our proposed method (with $f = 0.2$) with three baseline user profiles. The first type represents the interest of a user by a single set of the 10 most frequently used tags by the user. The second type is similar but includes the 20 most frequently used tags. The third type is in the form of multiple sets of tags like those generated by the proposed method, but the tags are randomly assigned to the sets. By using these baseline profiles, we aim to answer two questions: (1) Are the user profiles generated better than those single-set user profiles? (2) Does the cluster technique produce meaningful clusters for recommending interesting bookmarks to the users? The result of this experiment is plotted in Fig. 7(b).

Our results show that, when compared with the other baseline profiles, the profiles generated by the proposed method are able to retrieve more relevant bookmarks at the same similarity threshold. In other words, the user profiles allow a system to make better judgement regarding the relevance of a bookmark to the interests of a user. This suggests that the proposed method is able to break down a personomy into different meaningful sets of tags, so that a potentially interested bookmark can

be matched with a particular interest of the user more effectively. On the other hand, single-set user profiles (Top 10 and Top 20) which pool all tags together are likely to miss some bookmarks which are relevant to a specific interest of the user, and it does not help even when more tags are included in the profiles. This weakness is actually exacerbated when more tags are included in such type of user profiles, as we can see in the recall levels when user profiles with the top 20 tags are used.

In addition, Fig. 7(b) also shows that the user profiles generated by the proposed method perform significantly better than the randomly generated profiles. This suggests the clusters discovered by the proposed method are meaningful and truly reflect the diversity of the interests of the users.

5.3 Potential Applications

Our proposed algorithm provides a new way for constructing better user profiles based on the data available from collaborative tagging. There are a number of areas in which such algorithms can be employed.

Firstly, as the user profiles provide a summary of the different interests of the users, it can be readily used to facilitate the management and organisation of personal Web resources. In addition, the user profiles can also be used in Web page recommendation. Currently, Delicious provides various methods which allow users to keep track of new bookmarks which they may find interesting such as by subscribing to the RSS feed of a tag. However, there are currently no mechanisms which directly recommend interesting bookmarks to users. With the user profiles constructed by our proposed method, recommender systems will be able to recommend more specific bookmarks to users by targeting a particular interest of the users.

The proposed method of generating user profiles from folksonomies can easily be extended to accommodate other desirable features in user profiling. For example, by weighting the different interests with the number of bookmarks in the corresponding clusters, we are able to differentiate the major interests from other minor interests of a user. In addition, since the time at which a bookmark is saved can easily be obtained from the collaborative tagging system, it is also possible to determine whether an interest is a short-term or a long-term one of a user. We plan to investigate these features in our future work.

6 Related Work

User profile representation and construction has been a key research area in the context of personal information agents and recommendation systems. The representation of user profiles concerns with how user interests and preferences are modelled in a structured way. Probably the simplest form of user profile is a term vector indicating which terms are interested by the user. The weights in the vector is usually determined by the *tf-idf* weighting scheme as terms are extracted from documents

interested by the user or obtained by observing user behaviour [3, 15]. More sophisticated representations such as the use of a weighted network of n-grams [26] have also been proposed. However, a single user profile vector may not be enough when users have multiple interests in diverse areas [7], and several projects have employed multiple vectors to represent a user profile. For example, Pon et al. [23] use multiple profile vectors to represent user interests to assist recommendation of news articles. Kook [13] also proposes a Web user agent which represents user interests in multiple domains using multiple interest vectors.

In recent years, user-profiling approaches utilising the knowledge contained in ontologies have been proposed. In these approaches, a user profile is represented in terms of the concepts in an ontology which the user is interested in. For example, Middleton et al. [18] propose two experimental systems in which user profiles are represented in terms of a research paper topic ontology. Similar approaches have also been proposed to construct user profiles for assisting Web search [30] or enhancing recommendations made by collaborative filtering systems [2].

On the other hand, since the rise in popularity of collaborative tagging systems, some studies have also focused on generating user profiles from folksonomies. For example, in [5] a user profile is represented in the form of a tag vector, in which each element in the vector indicates the number of times a tag has been assigned to a document by the user. In [17], three different methods for constructing user profiles out of folksonomy data have been proposed. The first and simplest approach is to select the top k mostly used tags by a user as his profile. The second approach involves constructing a weighted network of co-occurrence of tags and selecting the top k pairs of tags which are connected by the edges with largest weights. The third method is an adaptive approach called the *Add-A-Tag* algorithm, which takes into account the time-based nature of tagging by reducing the weights on edges connecting two tags as time passes. In addition, Li et al. [14] introduce ISID, the Internet Social Interest Discovery system, which performs large scale clustering on tags and documents to group documents of similar topics together, thus finding out the common interests of the user community.

On the other hand, [22] discusses the issue of constructing a user profile from a folksonomy in the context of personalised Web search. In their approach, a user profile p_u is represented in the form of a weighted vector with m components (corresponding to the m tags used by the user). The use of $\boldsymbol{w}_d$ is to assign a weight between 0 and 1 to each of the n documents. While these attempts provide some possible methods for constructing user profiles based on data in folksonomies, the possibility of a user having multiple interests is not explicitly addressed in these works.

7 Conclusions

The emergence of collaborative tagging systems provide valuable sources of information for understanding user interests and constructing better user profiles. In this chapter, we investigate the characteristics of personomies extracted from

folksonomies, and observe that the interests of many users are very diverse, and they cannot be modelled by simple methods such as a single set of tags. A novel method for constructing user profiles which take into account the diversity of interests of the users is proposed. We evaluate the user profiles by looking at whether they provide a good summary of the bookmarks of the users. We also show that the user profiles generated by using our method are able to provide better support to recommender systems.

We believe that this research work provides valuable insight into how user profiles of multiple interests can be constructed out of a folksonomy. From this point, we plan to carry out further research work in three main directions. Firstly, we will further investigate how the proposed method can be improved. In our study, a user profiles constructed treats every cluster of bookmarks and its signature as corresponding to a distinctive interest of the user. However, it may be true that two interests are related and are only sub-topics of a more general area. We will investigate if the introduction of a hierarchical structure is desirable. Secondly, we will investigate whether the introduction of weights of tags would help improve the usefulness of the generated user profiles. While the occurrence frequency of the tags is considered in this work, associating weights with the tags in the clusters may facilitate better matching between the signatures and relevant resources. Furthermore, we will also attempt to extend our method to accommodate features such as the relative importance of different interests and the differentiation between long-term and short-term interests. We hope this research will ultimately deliver useful algorithms and applications which utilise the power of user-contributed metadata in folksonomies.

Acknowledgments We would like to thank the Drs Richard Charles and Esther Yewpick Lee Charitable Foundation which has awarded the R C Lee Centenary Scholarship to the first author of this paper to support his doctoral study at the University of Southampton, where the research work described in this paper was carried out. We also thank the reviewers of this paper for their invaluable comments and suggestions.

References

1. Adomavicius, G., Tuzhilin, A.: Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions. IEEE Trans. on Knowl. and Data Eng. **17**(6), 734–749 (2005)
2. Anand, S.S., Kearney, P., Shapcott, M.: Generating semantically enriched user profiles for web personalization. ACM Trans. Inter. Tech. **7**(4), 22 (2007)
3. Balabanovic, M., Shoham, Y.: Learning information retrieval agents: Experiments with automated web browsing. In: Proceedings of the AAAI Spring Symposium on Information Gathering from Heterogenous, Distributed Resources, March, 1995, Stanford, CA, USA, pp. 13–18 (1995)
4. Chirita, P.A., Damian, A., Nejdl, W., Siberski, W.: Search strategies for scientific collaboration networks. In: P2PIR '05: Proceedings of the 2005 ACM workshop on Information retrieval in peer-to-peer networks, pp. 33–40. ACM Press, New York, NY, USA (2005)

5. Diederich, J., Iofciu, T.: Finding communities of practice from user profiles based on folksonomies. In: Proceedings of the 1st International Workshop on Building Technology Enhanced Learning solutions for Communities of Practice (2006)
6. Girvan, M., Newman, M.E.J.: Community structure in social and biological networks. PROC.NATL.ACAD.SCI.USA **99**, 7821 (2002)
7. Godoy, D., Amandi, A.: User profiling in personal information agents: a survey. Knowl. Eng. Rev. **20**(4), 329–361 (2005)
8. Golder, S., Huberman, B.A.: Usage patterns of collaborative tagging systems. J. Inf. Sci. **32**(2), 198–208 (2006)
9. Grcar, M., Mladenić, D., Grobelnik, M.: User profiling for interest-focused browsing history. In: SIKDD 2005 at Multiconference IS 2005. Ljubljana, Slovenia (2005)
10. Guimera, R., Amaral, L.A.N.: Functional cartography of complex metabolic networks. Nature **433**, 895 (2005)
11. Hotho, A., Jäschke, R., Schmitz, C., Stumme, G.: Bibsonomy: A social bookmark and publication sharing system. In: A. de Moor, S. Polovina, H. Delugach (eds.) Proceedings of the Conceptual Structures Tool Interoperability Workshop at the 14th International Conference on Conceptual Structures. Aalborg University Press, Aalborg, Denmark (2006)
12. Hotho, A., Jäschke, R., Schmitz, C., Stumme, G.: Information retrieval in folksonomies: Search and ranking. In: Y. Sure, J. Domingue (eds.) The Semantic Web: Research and Applications, *LNCS*, vol. 4011, pp. 411–426. Springer Berlin (2006)
13. Kook, H.J.: Profiling multiple domains of user interests and using them for personalized web support. In: D.S. Huang, X.P. Zhang, G.B. Huang (eds.) Advances in Intelligent Computing, Proceedings of International Conference on Intelligent Computing (ICIC 2005), Part II, 23-26 August, 2005, Hefei, China, pp. 512–520. Springer-Verlag New York, Inc., Secaucus, NJ, USA (2005)
14. Li, X., Guo, L., Zhao, Y.E.: Tag-based social interest discovery. In: WWW '08: Proceeding of the 17th international conference on World Wide Web, Beijing, China, April 21–25, 2008, pp. 675–684. ACM, New York, NY, USA (2008)
15. Lieberman, H.: Letizia: An agent that assists web browsing. In: C.S. Mellish (ed.) Proceedings of the Fourteenth International Joint Conference on Artificial Intelligence (IJCAI-95), 20-25 August, 1995, Montreal, Quebec, Canada, pp. 924–929. Morgan Kaufmann Publishers Inc., San Mateo, CA, USA (1995)
16. Marlow, C., Naaman, M., Boyd, D., Davis, M.: Ht06, tagging paper, taxonomy, flickr, academic article, to read. In: HYPERTEXT '06: Proceedings of the Seventeenth Conference on Hypertext and Hypermedia, pp. 31–40. New York, NY, USA (2006)
17. Michlmayr, E., Cayzer, S.: Learning user profiles from tagging data and leveraging them for personal(ized) information access. In: Proceedings of the Workshop on Tagging and Metadata for Social Information Organization, Co-located with the 16th International World Wide Web Conference (WWW2007), 8–12 May, 2007, Banff, Alberta, Canada (2007)
18. Middleton, S.E., Shadbolt, N.R., Roure, D.C.D.: Ontological user profiling in recommender systems. ACM Trans. Inf. Syst. **22**(1), 54–88 (2004)
19. Mika, P.: Ontologies are us: A unified model of social networks and semantics. J. Web Semant. **5**(1), 5–15 (2007)
20. Newman, M.E.J.: Fast algorithm for detecting community structure in networks. Phy. Rev. E **69**, 066,133 (2004)
21. Newman, M.E.J., Girvan, M.: Finding and evaluating community structure in networks. Phy. Rev. E **69**, 026,113 (2004)
22. Noll, M., Meinel, C.: Web search personalization via social bookmarking and tagging. In: K. Aberer, K.S. Choi, N.F. Noy, D. Allemang, K.I. Lee, L.J.B. Nixon, J. Golbeck, P. Mika, D. Maynard, R. Mizoguchi, G. Schreiber, P. Cudré-Mauroux (eds.) Proceedings of the 6th International Semantic Web Conference and 2nd Asian Semantic Web Conference (ISWC/ASWC2007), LNCS 4825, Busan, South Korea, 11–15 November, 2007, pp. 365–378. Springer-Verlag, Berlin, Germany (2007)

23. Pon, R.K., Cardenas, A.F., Buttler, D., Critchlow, T.: Tracking multiple topics for finding interesting articles. In: KDD '07: Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 560–569. ACM, New York, NY, USA (2007)
24. Radicchi, F., Castellano, C., Cecconi, F., Loreto, V., Parisi, D.: Defining and identifying communities in networks. PROC.NATL.ACAD.SCI.USA **101**, 2658 (2004)
25. van Rijsbergen, C.J.: Information Retrieval. Dept. of Computer Science, University of Glasgow (1979)
26. Sorensen, H., Mcelligot, M.: Psun: A profiling system for usenet news. In: CKIM'95 Workshop on Intelligent Information Agents (1995)
27. Sugiyama, K., Hatano, K., Yoshikawa, M.: Adaptive web search based on user profile constructed without any effort from users. In: WWW '04: Proceedings of the 13th International Conference on World Wide Web, pp. 675–684. ACM Press, New York, NY, USA (2004)
28. Vander Wal, T.: Folksonomy definition and wikipedia. *http://www.vanderwal.net/random /entrysel.php?blog=1750*, November 2, 2005. Accessed 13 Feb 2008.
29. Wasserman, S., Faust, K.: Social Network Analysis. Cambridge University Press, Cambridge (1994)
30. Zhou, X., Wu, S.T., Li, Y., Xu, Y., Lau, R.Y.K., Bruza, P.D.: Utilizing search intent in topic ontology-based user profile for web mining. In: WI '06: Proceedings of the 2006 IEEE/WIC/ACM International Conference on Web Intelligence, pp. 558–564. IEEE Computer Society, Washington, DC, USA (2006)