A Study of User Profile Generation from Folksonomies

Ching-man Au Yeung
cmay06r@ecs.soton.ac.uk

Nicholas Gibbins
nmg@ecs.soton.ac.uk

Nigel Shadbolt
nrs@ecs.soton.ac.uk

Intelligence, Agents, Multimedia Group
School of Electronics and Computer Science
University of Southampton
Southampton, SO17 1BJ, United Kingdom

ABSTRACT

Recommendation systems which aim at providing relevant information to users are becoming more and more important and desirable due to the enormous amount of information available on the Web. Crucial to the performance of a recommendation system is the accuracy of the user profiles used to represent the interests of the users. In recent years, popular collaborative tagging systems such as del.icio.us have aggregated an abundant amount of user-contributed metadata which provides valuable information about the interests of the users. In this paper, we present our analysis on the personal data in folksonomies, and investigate how accurate user profiles can be generated from this data. We reveal that the majority of users possess multiple interests, and propose an algorithm to generate user profiles which can accurately represent these multiple interests. We also discuss how these user profiles can be used for recommending Web pages and organising personal data.

Categories and Subject Descriptors

H.3.4 [**Information Storage and Retrieval**]: Systems and Software; H.3.5 [**Information Storage and Retrieval**]: Online Information Services; H.5 [**Information Interfaces and Presentation (I.7)**]:

General Terms

Algorithms, Experimentation, Human Factors

Keywords

collaborative tagging, folksonomy, personomy, user profile

1. INTRODUCTION

The amount of resources on the Web nowadays is so enormous that retrieval of relevant information is getting more and more difficult. While users are desperate to obtain information that is relevant to their needs and to avoid information that are irrelevant, publishers of resources are also eager to deliver their information to their targeted readers. This has resulted in the rise of recommendation systems [3] which aim to recommend relevant and interesting resources to users. An important aspect of user profiles is whether they can truly reflect the interests or expertise of the users. Some research works attempt to construct user profiles based on the browsing history of the users [9, 22], or on the documents collected by the users [4].

Recently, the rising popularity of collaborative tagging systems [8], such as del.icio.us[1] and Flickr[2], has provided new sources for understanding the interests of Web users. Collaborative tagging systems allow users to choose their own words as tags to describe their favourite Web resources, resulting in an emerging classification scheme now commonly known as a *folksonomy* [24]. Given that the resources and the tags posted by Web users to these systems are supposed to be highly dependent on their interests, folksonomies thus provide rich information for building more accurate and more specific user profiles for use in various applications.

Currently, only a few studies in the literature try to construct user profiles from data in collaborative tagging systems [5, 13], and usually only a single set of popular tags are used to represent user interests. However, we observe that tags used by users are very diverse and span across many different domains. This implies that users usually have a wide range of interests. Therefore, a single set of tags may not be the most suitable representation of a user profile, as it is not able to reflect the multiple interests of users. In this paper, we propose a network analysis technique performed on the personomy [11] of a user to identify the different interests of a user, and to construct a more comprehensive user profile based on the results. Evaluations show that our algorithm is able to reveal the different domains in which the users are interested, and construct more informative and specific user profiles.

This paper is structured as follows. Section 2 introduces folksonomies and personomies. Section 3, presents the analysis of the data collected from del.icio.us which motivated this research. In Section 4, we describe in detail our proposed algorithm for user profile construction. Evaluations, discussions and potential applications are presented in Section 5. We mentioned related works in Section 6. Finally, Section 7 concludes the paper and gives future research directions.

2. FOLKSONOMIES AND PERSONOMIES

Folksonomies [24] are user-contributed data aggregated by collaborative tagging systems. In these systems, users are allowed to choose terms freely to describe their favourite Web resources. A folksonomy is generally considered to consist

Copyright is held by the Authors. Copyright transfered for publishing online and a conference CD ROM.
SWKM'2008: Workshop on Social Web and Knowledge Management @ WWW 2008, April 22, 2008, Beijing, China.

[1] http://del.icio.us/
[2] http://www.flickr.com/

of at least three sets of elements, namely users, tags and resources. Although there can be different kinds of resources, in this article we will focus on Web documents, such as those being bookmarked in del.icio.us. Formally, a folksonomy is defined as follows [15].

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.

If we want to understand the interests of a single user, we only need to concentrate on the tags and documents that are associated with this particular user. Such set of data is given the name ***personomy*** [11].[3]

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: $D_u = \{d|(t, d) \in A_u\}$.

This definition is identical to the one mentioned in [11], except that we choose to exclude the sub-tag/super-tag relation, since most collaborative tagging systems do not offer such functionality and we will not deal with this here.

To perform analysis on the personomy of a user, we first represent the personomy in the form of a network, with nodes representing tags and documents associated with the 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 by extracting the part that is related to the user. The bipartite graph TD_u of a personomy of a user u is defined as follows.

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

An edge exists between a tag and a document if the tag is assigned to the document. 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.

To perform document clustering, we can fold the bipartite graph into a one-mode network [15] of documents: $\mathbf{D} = \mathbf{X}'\mathbf{X}$. The adjacency matrix $\mathbf{D}$ represents the personal repository of the user. Links between documents are weighted by the number of tags that have been assigned to both documents. Thus, documents with higher weights on the links between them can be considered as more related. On the other hand, a one-mode network of tags can be constructed in a similar fashion: $\mathbf{T} = \mathbf{X}'\mathbf{X}$. $\mathbf{T}$ represents semantic network which shows the associations between different tags. In other words, this is the personal vocabulary or a simple ontology used by the particular user.

To facilitate the following discussions, we further define several notations here. Firstly, we denote the set of documents tagged by the tag t in the personomy of user u by $D_{u,t}$:

$$D_{u,t} = \{d|(t, d) \in A_u\}$$

Also, we define $Co_u(t_1, t_2)$ which indicates whether two tags t_1 and t_2 have been used on the same document by a user:

$$Co_u(t_1, t_2) = \begin{cases} 1 & \text{if } (t_1, d) \in A_u,\ (t_2, d) \in A_u \text{ for some } d \\ 0 & \text{otherwise} \end{cases}$$

[3]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/.

Total number of users		9,185
Tags	Maximum	18,952
	Minimum	1
	Mean	285
Bookmarks	Maximum	34,201
	Minimum	1
	Mean	602

Table 1: Summary of data obtained from del.icio.us.

3. ANALYSIS OF PERSONOMIES

To understand the characteristics of personomies in collaborative tagging systems, we perform analysis on data collected from del.icio.us. In particular, we want to gain insight into the general behaviour of Web users using these systems. We also want to understand if users are generally interested in a rather specific domain, such as we might expect when studying the publications of a researcher, or if they are more likely to be interested in a wide range of topics.

In December 2007, we collected the bookmarking data of 9,431 users of del.icio.us, including their bookmarks and the tags they used, by crawling del.icio.us user names which appeared on the page showing the recently updated bookmarks.[4] It is noted that among the 9,431 users whose data we have collected, 246 of them apply no tags to any of their stored bookmarks. These users are filtered when performing the following analysis. We summarise the statistics of the data of the remaining 9,185 users in Table 1 and Figure 1.

3.1 Number of Tags and Bookmarks of a User

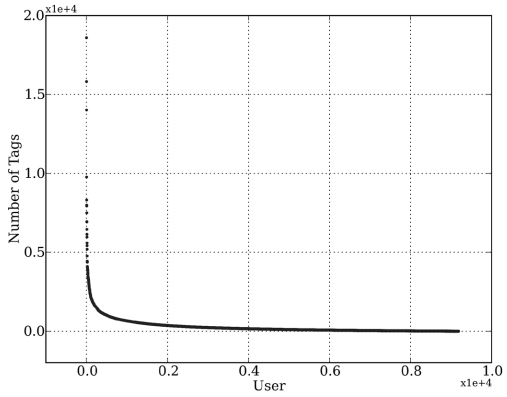
From the summary of the data in Table 1 and Figure 1, we can see that on average a user have used 285 unique tags and have saved 602 unique bookmarks on del.icio.us. Although some users have over 18,000 tags and over 34,000 bookmarks, only a very small number of users have more than a thousand tags or bookmarks. This finding agrees with what Golder and Huberman [8] report in their paper, 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, suggesting a power-law distribution.

In addition, 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 shows a moderate relationship between the number of tags and the number of bookmarks, with a correlation coefficient of 0.55.

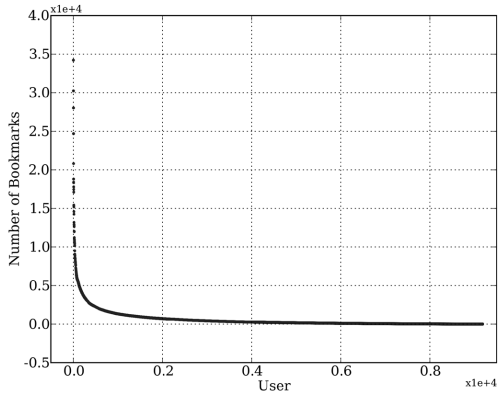
In fact, it is natural to suggest that when there are more bookmarks more tags are required to distinguish between different bookmarks by putting them into more specific categories. However the bookmarks and tags of the users in the system are also highly dependent on the interests of the users. If a user has a very specific interest, a small number of tags will be enough for even a large number of bookmarks, as 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 is stronger for users with fewer bookmarks than those with many bookmarks. For users with fewer than 500 bookmarks, the correlation coefficient is 0.43. For users with more than 5,000 bookmarks, the

[4]http://del.icio.us/recent



(a) Tags



(b) Bookmarks

Figure 1: Number of tags and bookmarks of the users.

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 Multiple Interests of Users

With the average number of bookmarks significantly larger than the average number of tags being used, it is obvious that users are very likely to use a tag to describe more than one bookmark. However, the usage of tags also depends on the diversity of interests of the users. A user with only one or two specific interests is likely to use fewer tags than another user who is interested in topics across several different domains. To understand this aspect of users in collaborative tagging system, we propose two measures which reflect the diversity of interests of the users. We will give examples based on the two fictional users in Table 2, one with rather specific interests in Semantic Web related topics, while another has more diverse interests such as cooking and sports.

Firstly, we study the relations between the tags and the bookmarks. If the tags used by a user are all assigned to most of the bookmarks, the user is likely to have a rather specific interest, because this set of tags applies to most of

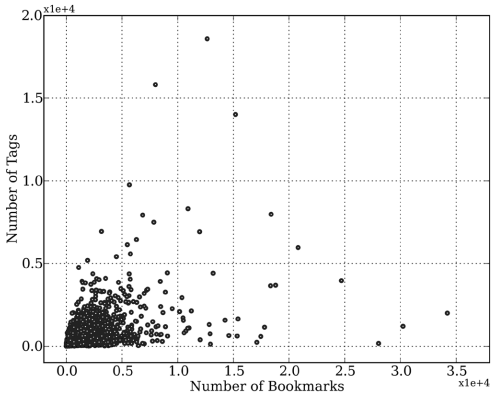


Figure 2: Scatter plot of number of tags against number of bookmarks.

user	bookmark	tags
u_1	d_1	web2.0, semanticweb, ontology, notes
	d_2	semanticweb, ontology
	d_3	semanticweb, ontology, RDF
u_2	d_4	semanticweb, folksonomy, tagging
	d_5	toread, cooking, recipe, food
	d_6	sports, football, news

Table 2: Two example users with their personomies.

the documents that the user is interested in. On the other hand, if most of the tags are only used on a small fraction of bookmarks, it is likely that the user has a broader range of interests. To quantify this characteristic, we propose a measure called ***tag utilisation*** which is defined as follows.

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

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

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 so the user should have 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 should have multiple interests which are quite distinctive from each other. Such characteristic can be measure by ***average tag co-occurrence ratio***, which is defined as follows.

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

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

If we represent the co-occurrences between the tags as a network (by constructing the adjacency matrix $\mathbf{T}$), we can

	MAX	MIN	MEAN	STD
TU	1.0000	0.0003	0.0617	0.1388
ATCR	1.0000	0.0000	0.0707	0.1297

Table 3: Summary of the two measures of the data.

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:

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

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 [25] 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 2. 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 bookmark collection.

Next, we apply these two measures on the set of data that we have collected from del.icio.us. The results are summarised in Table 3 and Figure 3.

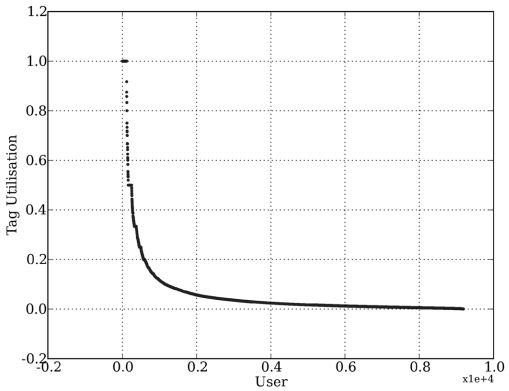
Although the two measures are designed to measure different characteristics of personomies, the results do have very common features. Firstly, the mean values of tag utilisation and average tag co-occurrence ratio both very low, at 0.06 and 0.07 respectively, even though the values span across the whole range from 0 to 1. These values mean 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 Figure 3 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 of personomies have values less than 0.2 (93% in both measures). Also, there is a strong correlation between tag utilisation and average tag co-occurrence ratio, with a correlation coefficient of 0.71.

Given these figures, we reveal 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 suggests that the bookmarks of the users have topics which are rather diverse such that tags do not apply to all of them. Also, a user's tags can be terms from different domains which are not used together very often on the bookmarks. Hence, this indicates that users of del.icio.us have diverse interests instead of a single interest in a very specific domains.

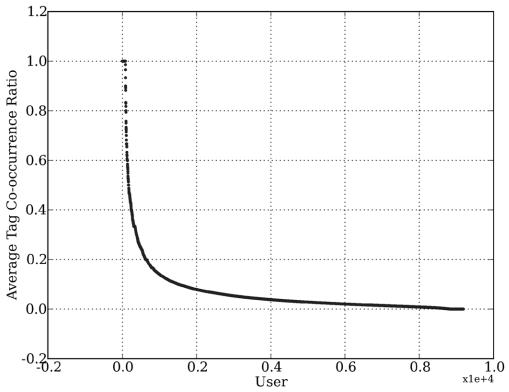
4. USER PROFILE CONSTRUCTION

As the majority of users in del.icio.us 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, user profiles which can accommodate the multiple interests of the users are very much desirable.

Identifying the different interests can be a challenging task



(a) Tag utilisation



(b) Average tag co-occurrence ratio

Figure 3: Distribution of tag utilisation and average tag co-occurrence. ratio

as tags are freely chosen by users and their actual meaning is usually not very 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 tagged by similar tags, we can perform clustering algorithms on the documents tagged by a user to group documents of similar topics together, and extract the sets of tags assigned to these documents as indicators of the users' different interests.

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 reflect the interests of the users.

4.1 Community Discovery Algorithms

Clusters in a network are basically 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 [20].

To quantitatively measure the 'goodness' of the clusters discovered, the measure of ***modularity*** [17] 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 to optimise 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 [17]. 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 [16]. In this paper, we will employ this fast greedy algorithm to perform clustering, as it is efficient and performs well on large networks.

4.2 Construction of User Profiles

Given a network of documents (which are bookmarks in our case), we can 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 as corresponding 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 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 $K_i's$: $P_u = \{K_i\}$.

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 4 shows the results of applying the proposed method on two personomies, with $f = 20\%$.

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

User B	
K_1	webdesign, programming
K_2	interesting, art, video, funny
K_3	food, books, tobuy
K_4	lort, debate

Table 4: User profile constructed for two users.

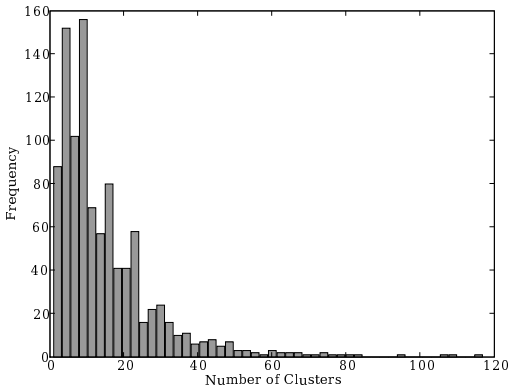


Figure 4: Number of clusters discovered for the 1000 personomies.

5. EVALUATION AND DISCUSSIONS

From our data set, we select at random 1,000 users 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 4 graphs the number of clusters discovered for each of the personomies. On average 15 clusters are discovered in each personomy.

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. Therefore we try to evaluate our proposed method by asking the following question: are the sets of tags accurate descriptions of the clusters of bookmarks from which they are extracted? If this is the case, then the user profiles should accurately represent the interests of the users. In the following we present the evaluations which

attempt to answer this question.

5.1 Precision and Recall Measures

Our 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 described a cluster? To answer such questions, we will employ the measures of precision and recall [23] 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-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 $\vec{e} = (e_1, e_2, ..., e_n)$ 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 $\vec{q}$. The retrieval process is carried out by calculating the cosine similarity between the query vector and the bookmark vectors:

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

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 [23] which is a combined measure of precision and recall.

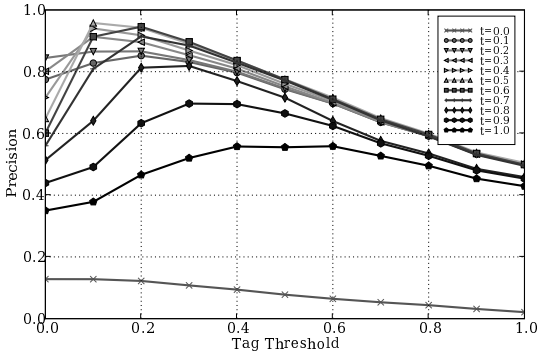
$$Precision(c) = \frac{|D_x \cap D_c|}{|D_x|} \quad (5)$$

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

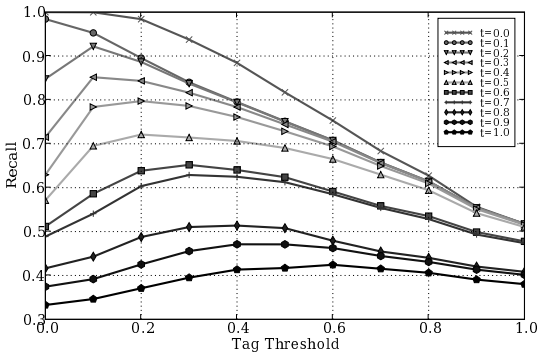
$$F_1(c) = \frac{2 \times Precision(c) \times Recall(c)}{Precision(c) + Recall(c)} \quad (7)$$

We calculated the three measures for the user profiles generated from the 1,000 selected personomies. We control two parameters in our evaluation, one is the value of f, the percentage of bookmarks above which a tag is assigned to in a cluster for it to be included in the signature, and the value of t, the threshold of cosine similarity. The results are presented in Figure 5.

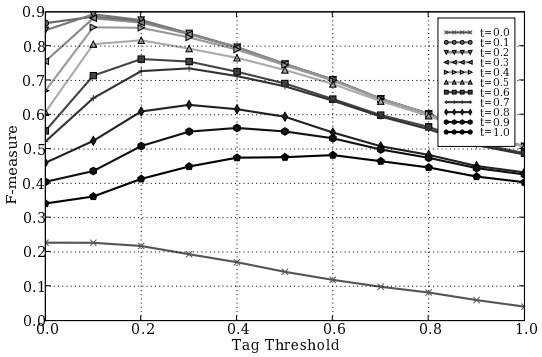
Figure 5(a) shows that for most values of similarity threshold precision attains maximum for f in the range from 0.1 to 0.4, and thereafter it continues to decrease as f increases. The result suggests that if only the most common tags are included in the signatures, they will become less representative as summaries of the clusters. This is probably due to the fact that the most common tags are usually too general



(a) Precision



(b) Recall



(c) F_1 measure

Figure 5: Precision, recall and F_1 measure. Different lines correspond to different values of similarity threshold.

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 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 $t = 0.3$ to $t = 0.5$, 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 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 applied to. We briefly discuss two of them in this section.

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 organization of personal Web resources. For example, the sets of tags representing the clusters of bookmarks can be used to facilitate navigation and retrieval of a user's own bookmarks in del.icio.us. This would be much more efficient than navigating through the bookmarks by a single tag.

In addition, the user profiles can also be used to support Web page recommendation systems. Currently, del.icio.us provides various methods which allow users to keep track of new bookmarks which they may find interesting, such as subscribing to the RSS feed of a tag, or adding a user of similar interests to one's network. However, there have been no mechanisms which directly recommend interesting bookmarks to the users. With the user profiles constructed by our proposed method, recommendation systems will have a better understanding of the interests of the users, and be able to recommend more specific bookmarks to users by targeting a particular interest of the users.

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 [2, 12]. More sophisticated representations such as the use of a weighted network of n-grams [21] 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. [19] use multiple profile vectors to represent user interests to assist recommendation of news articles. In recent years, user-profiling approaches utilizing the knowledge contained in ontologies have been proposed. In these approaches, a user profile is represented in terms of the concepts that the user is interested in an ontology. For example, Middleton et al. [14] 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 searching [26] or enhancing recommendations from collaborative filtering systems [1].

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 generator based on the annotations assigned by the users to the documents is proposed. The user profile is represented in the form of a tag vector, which each element in the vector indicating the number of times a tag has been assigned to a document by the user. In [13], 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, [18] 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 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 paper, we investigated the characteristics of personomies extracted from folksonomies, and observed that the majority of users possess a wide range of interests, which 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 also evaluated the user profiles by looking at whether they provide a good summary of the bookmarks of the users.

This research work provides insight into how user profiles of multiple interests can be constructed based on the data collected from a folksonomy. From this point, we plan to carry out further research work in two 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 cor-

responding to a distinctive interest of the user. However, it may be true that two interests are related and are only subtopics of a more general area. We will investigate if the introduction of a hierarchical structure is desirable. Secondly, we will attempt to evaluate our proposed method by applying the user profiles on applications such as Web page recommendation or personal resource management. We hope this research will ultimately deliver useful algorithms and applications which utilise the power of user-contributed metadata in collaborative tagging systems.

8. REFERENCES

[1] Sarabjot Singh Anand, Patricia Kearney, and Mary Shapcott. Generating semantically enriched user profiles for web personalization. *ACM Trans. Inter. Tech.*, 7(4):22, 2007.

[2] Marko Balabanovic and Yoav Shoham. Learning information retrieval agents: Experiments with automated web browsing. In *Proceedings of the AAAI Spring Symposium on Information Gathering from Heterogenous, Distributed Resources*, pages 13–18, 1995.

[3] Robin Burke. Hybrid recommender systems: Survey and experiments. *User Modeling and User-Adapted Interaction*, 12(4):331–370, 2002.

[4] Paul Alexandru Chirita, Andrei Damian, Wolfgang Nejdl, and Wolf Siberski. Search strategies for scientific collaboration networks. In *P2PIR '05: Proceedings of the 2005 ACM workshop on Information retrieval in peer-to-peer networks*, pages 33–40, New York, NY, USA, 2005. ACM Press.

[5] Jörg Diederich and Tereza Iofciu. 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] M. Girvan and M. E. J. Newman. Community structure in social and biological networks. *PROC.NATL.ACAD.SCI.USA*, 99:7821, 2002.

[7] Daniela Godoy and Analia Amandi. User profiling in personal information agents: a survey. *Knowl. Eng. Rev.*, 20(4):329–361, 2005.

[8] Scott Golder and Bernardo A. Huberman. Usage patterns of collaborative tagging systems. *Journal of Information Science*, 32(2):198–208, April 2006.

[9] Miha Grcar, Dunja Mladenić, and Marko Grobelnik. User profiling for interest-focused browsing history. In *SIKDD 2005 at Multiconference IS 2005*, Ljubljana, Slovenia, 2005.

[10] Roger Guimera and Luis A. Nunes Amaral. Functional cartography of complex metabolic networks. *Nature*, 433:895, 2005.

[11] Andreas Hotho, Robert Jäschke, Christoph Schmitz, and Gerd Stumme. Information retrieval in folksonomies: Search and ranking. In York Sure and John Domingue, editors, *The Semantic Web: Research and Applications*, volume 4011 of *LNCS*, pages 411–426. Springer, June 2006.

[12] Henry Lieberman. Letizia: An agent that assists web browsing. In Chris S. Mellish, editor, *Proceedings of the Fourteenth International Joint Conference on Artificial Intelligence (IJCAI-95)*, pages 924–929, Montreal, Quebec, Canada, 1995. Morgan Kaufmann publishers Inc.: San Mateo, CA, USA.

[13] Elke Michlmayr and Steve Cayzer. 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, 16th International World Wide Web Conference (WWW2007)*, May 2007.

[14] Stuart E. Middleton, Nigel R. Shadbolt, and David C. De Roure. Ontological user profiling in recommender systems. *ACM Trans. Inf. Syst.*, 22(1):54–88, 2004.

[15] Peter Mika. Ontologies are us: A unified model of social networks and semantics. *Web Semant.*, 5(1):5–15, 2007.

[16] M. E. J. Newman. Fast algorithm for detecting community structure in networks. *Physical Review E*, 69:066133, 2004.

[17] M. E. J. Newman and M. Girvan. Finding and evaluating community structure in networks. *Physical Review E*, 69:026113, 2004.

[18] Michael Noll and Christoph Meinel. Web search personalization via social bookmarking and tagging. In *Proceedings of the 6th International Semantic Web Conference and 2nd Asian Semantic Web Conference (ISWC/ASWC2007), Busan, South Korea*, pages 365–378, November 2007.

[19] Raymond K. Pon, Alfonso F. Cardenas, David Buttler, and Terence Critchlow. Tracking multiple topics for finding interesting articles. In *KDD '07: Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 560–569, New York, NY, USA, 2007. ACM.

[20] Filippo Radicchi, Claudio Castellano, Federico Cecconi, Vittorio Loreto, and Domenico Parisi. Defining and identifying communities in networks. *PROC.NATL.ACAD.SCI.USA*, 101:2658, 2004.

[21] H. Sorensen and M. Mcelligot. Psun: A profiling system for usenet news. In *CKIM'95 Workshop on Intelligent Information Agents*, 1995.

[22] Kazunari Sugiyama, Kenji Hatano, and Masatoshi Yoshikawa. 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*, pages 675–684, New York, NY, USA, 2004. ACM Press.

[23] C. J. van Rijsbergen. *Information Retrieval*. Dept. of Computer Science, University of Glasgow, 1979.

[24] Thomas Vander Wal. Folksonomy definition and wikipedia. *http://www.vanderwal.net/random/entrysel.php?blog=1750*, November 2, 2005. Accessed 13 Feb 2008.

[25] S. Wasserman and K. Faust. *Social network analysis*. Cambridge University Press, Cambridge, 1994.

[26] Xujuan Zhou, Sheng-Tang Wu, Yuefeng Li, Yue Xu, Raymond Y. K. Lau, and Peter D. Bruza. 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*, pages 558–564, Washington, DC, USA, 2006. IEEE Computer Society.