Discovering and Modelling Multiple Interests of Users in Collaborative Tagging Systems

Ching-man Au Yeung Nicholas Gibbins Nigel Shadbolt
Intelligence, Agents, Multimedia Group
School of Electronics and Computer Science
University of Southampton
Southampton, SO17 1BJ, UK
{cmay06r,nmg,nrs}@ecs.soton.ac.uk

Abstract

We analyse data obtained from several collaborative tagging systems and discover that user interests 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 using data in a collaborative tagging system. Our evaluation suggests that the proposed method is able to generate user profiles which reflect the diversity of user interests and can be used to help provide more focused recommendation.

1 Introduction

Collaborative tagging systems such as Delicious have provided new opportunities for understanding the interests of Web users, which can be used to build better user profiles for recommender systems. There are only a few studies in the literature which try to construct user profiles based on the information available in these systems [2, 6], and usually only a single set of tags are used 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 may not be a 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 personomies [4] of the users to identify their multiple interests and to construct more comprehensive user profiles.

2 Folksonomies and Personomies

Folksonomies are user-contributed metadata in collaborative tagging systems. A folksonomy consist of three sets of elements, namely users, tags and documents, and is usually defined as 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.

We can extract from a folksonomy tags and documents which are associated with a particular user. Such set of data is given the name ***personomy*** [4]. A personomy $\mathbf{P}_u$ of a user u is defined by restricting 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: $D_u = \{d|(t, d) \in A_u\}$.

We represent a personomy in the form of a graph, with nodes representing the tags and documents associated with a 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: $G_u = \langle T_u \cup D_u, E\rangle$, where $E = \{(t, d)|(t, d) \in A_u\}$. An edge exists between a tag and a document if the tag has been assigned to the document. The graph can be represented in matrix form: $\mathbf{X} = \{x_{ij}\}$, and $x_{ij} = 1$ 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 [7] of documents: $\mathbf{A} = \mathbf{X}^{\mathrm{T}}\mathbf{X}$. The adjacency matrix $\mathbf{A}$ represents the personal repository of the user. Edges between two documents are weighted by the number of tags that have been assigned to both of them. Thus, documents with higher weights on the edges between them can be considered as more related.

3 Multiple Interests of Users

We propose two measures for studying the diversity of user interests. The first measure involves examining how frequently a tag is used on the resources of the users. In-

	users	tags	docs
Delicious	9,185	444,135	3,281,306
Bibsonomy	423	10,213	16,174
LibraryThing	8,531	405,468	1,665,339

Table 1. Summary of the data collected.

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 personomies.

tuitively, if a user is only interested in one or two topics, the tags used by this user should appear on most of the resources. To quantify this characteristic, we propose a measure called ***tag utilisation*** which is defined as follows.

Definition 1 *Tag utilisation of a user u is the average of the fractions of bookmarks on which a tag is used:*

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

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

In addition, diversity of user interests can also be understood by examining tag co-occurrence. If the tags of a user are always used together with each other, it is likely that the tags are about similar topics. As a result it can be suggested that the user has a rather specific interest. Such characteristic can be measure by the ***average tag co-occurrence ratio***.

Definition 2 *Average tag co-occurrence ratio of a user measures how likely two tags are used together on the same bookmark by a user:*

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

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 collection of resources.

In order to understand the diversity of user interests in collaborative tagging systems, we apply the two measures

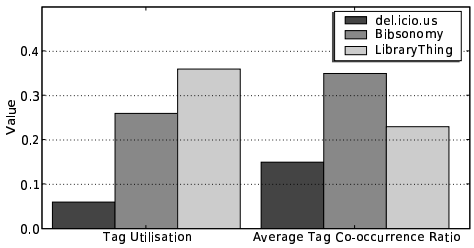

Figure 1. Tag utilisation and average tag co-occurrence ratio.

on personomies collected from three different collaborative tagging systems, namely Delicious, Bibsonomy and LibraryThing. We collect data using a crawler in the period between December 2007 and March 2008. For each of the systems, we crawl tags appearing on its front page and go on to collect data of users who have used these tags.

The average values of these two measures of the collected data are plotted in Fig. 1. The values of the two measures are less than 0.4 for all the three systems, with personomies in Delicious attaining lower values than those in LibraryThing and Bibsonomy. Bibsonomy is mainly for organising research publications, and the relatively high values in both measures probably reflect that users tend to be interested in only a few specific research topics. As for LibraryThing, we observe that tags are often used to classify books into broad categories such as 'fiction' or 'biography', and are seldom used to describe the content of the books. The relatively high values in the two measures probably reflect these facts.

On the other hand, Delicious can be used to organise any Web resources with an URL, thus it can be considered as the most general-purpose system among the three. Hence, it is reasonable that user interests in Delicious are most diverse. The average value of tag utilisation in Delicious is 0.06, meaning that a tag is on average only used on 6% of the bookmarks by a user. The average value of average co-occurrence ratio is 0.15, meaning that a tag is only used together with 15% of all tags used by a user. It can be concluded that users of collaborative tagging systems, especially those of social bookmarking sites such as Delicious, are mostly interested in a wide range of topics as indicated by their tag usage. Hence, it is essential that we have suitable methods which are able to extract the information and generate user profiles which truly reflect this diversity.

4 User Profile Generation

As the majority of users have diverse interests, user profiles which can accommodate the multiple interests of the

users are desirable. Firstly we need to first discover the multiple interests of the users. Given a network of documents constructed by using the method described in Section 2, we can employ clustering algorithms to group documents of similar topics together, extract the sets of tags assigned to these different group of documents, and use them to represent the multiple interests of the users. In this paper we use the fast-greedy community-discovery algorithm [8] to obtain a set of clusters of closely connected documents for a user. The algorithm is chosen because of its efficiency as well its good performance on a wide range of problems. We summarise our method for constructing a user profile for user u as follows.

1. Extract the personomy $\mathbf{P}_u$ of user u from the folksonomy $\mathbf{F}$, and construct the bipartite graph G_u.
2. Construct a one-mode network of documents out of G_u, and perform modularity optimization over the network of documents.
3. For each c_i of the n clusters obtained, extract a set K_i of tags which appear on more than $f\%$ of the documents as a signature of that cluster.
4. Return a user profile $P_u = \{K_i | i = 0, 1, ..., n\}$.

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 all of the bookmarks. Different choices will have different effect on the accuracy of the profile in representing the user's interest. We will investigate the problem of choosing a right value for f later. As an example, using this method we are able to extract two sets of tags for a user of Delicious, with one set having tags such as *webdesign*, *web2.0* and *css*, and another set having tags such as *linux*, *opensource* and *ubuntu*.

5 Evaluation and Discussion

We perform our evaluation on the personomies in Delicious in which interests are found to be most diverse. The data are divided into a training set and a test set. For each of the 9,185 users, we extract the first 70% bookmarks and the tags associated with them, and use them to generate a user profile using our proposed method. The generated user profile is then used to retrieve the remaining 30% of the bookmarks. The bookmarks are retrieved according to the similarity between the sets of tags in the user profile and the tags assigned to the bookmarks. We employ the following similarity measure between two sets of tags: $Sim(X, Y) = \frac{2\times|X\cap Y|}{|X|+|Y|}$.

The notion of recall in information retrieval research is adopted as a performance measure. It measures the fraction of relevant documents that can be retrieved by a certain

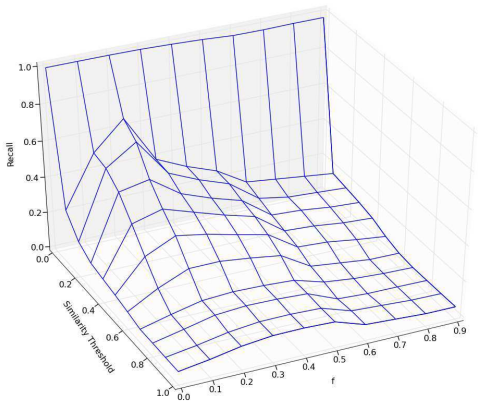

Figure 2. Recall at different values of f.

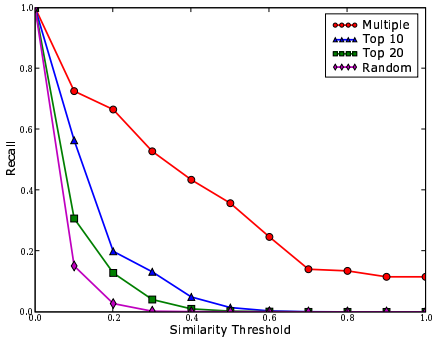

Figure 3. Recall of different types of profiles.

method. 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: $Recall(\alpha) = \frac{|D_i^\alpha \cap D_r|}{|D_i^\alpha|}$.

Our first experiment aims 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. Fig. 2 shows that the user profiles do not help retrieve relevant bookmarks when too few (large values of f) tags are included in the signatures of the clusters. The optimal value of $f \simeq 0.2$ means that more tags should be included in the signature for better recall.

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 and second types represent the interest of a user by a single set of the 10 or 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 with the tags randomly assigned to the sets. By using these baseline profiles, we aim at answering 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. 3.

Compare 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 generated user profiles allow a system to make better judgement regarding the relevance of a bookmark to 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 which pool all tags together are likely to miss some bookmarks which are relevant to a specific interest of the user. This weakness is actually exacerbated when more tags are included in such type of user profiles. In addition, we also see that the 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 user interests. Our evaluation thus gives positive answers to both questions we mentioned above.

6 Related Works

User profiling is a key research area in the context of recommender systems. The simplest form of user profile is a term vector indicating which terms are interested by the user [1]. More sophisticated methods involve the use of a weighted network of n-grams [9]. As single vector may not be enough when users have multiple interests [3], some projects employ multiple vectors to represent a user profile (e.g. [5]). There are also some studies which focus on generating user profiles from folksonomies. In [2] 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 used by the user. In [6], an adaptive approach is proposed which takes into account the time-based nature of tagging by reducing the weights on edges connecting two tags as time passes. These studies, however, do not explicitly address the possibility of a user having multiple interests.

7 Future Work and Conclusions

Collaborative tagging systems represent valuable sources of information for understanding user interests and constructing more accurate user profiles. We propose a method for constructing user profiles from folksonomies which take into account the diversity of interests of the users. In the future, we plan to extend our method to accommodate features such as the relative importance of different tags and interests with respect to a particular user.

References

[1] 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, March, 1995, Stanford, CA, USA*, pages 13–18, 1995.

[2] 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.

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

[4] 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.

[5] Hyung Joon Kook. Profiling multiple domains of user interests and using them for personalized web support. In De-Shuang Huang, Xiao-Ping Zhang, and Guang-Bin Huang, editors, *Proceedings of International Conference on Intelligent Computing, Part II, 23-26 August, 2005, Hefei, China*, pages 512–520. Springer-Verlag, 2005.

[6] 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, co-located with the 16th International World Wide Web Conference (WWW2007), 8-12 May, 2007, Banff, Alberta, Canada*, May 2007.

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

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

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