Collective User Behaviour and Tag Contextualisation in Folksonomies

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

Abstract

Collaborative tagging systems have emerged in recent years to become popular tools for organising information on the Web. While collaborative tagging offers many advantages, they also suffer from several limitations, with a major one being the existence of ambiguous tags. To understand what an ambiguous tag is intended to mean, we need to know the contexts in which it is used. Instead of using common large scale clustering techniques on folksonomies, we believe tags can be better contextualised by the social contexts in which they are used. We propose a method to reveal semantics of ambiguous tags by studying the collective user behaviour in a tagging system. In this paper we describe our proposal and some results of our preliminary experiments. We also discuss the significance of the work and how it can be evaluated.

1. Introduction

Collaborative tagging systems [7] such as Delicious[1] and Bibsonomy[2] have emerged in recent years to become popular tools for organising information on the Web. These systems allow Web users to use tags in the form of freely-chosen keywords to describe publicly available Web documents. The collaborative nature of these systems results in a continuously evolving categorisation scheme now commonly known as a folksonomy [20].

While collaborative tagging offers many advantages over the use of controlled vocabularies in categorising documents [6, 13], it also suffer from some limitations due to its unrestricted nature [7]. In particular, as users are free to use any terms or phrases as tags, a tag can be used to refer to several different concepts depending on the contexts in which it is used [7, 21]. The fact that many tags are ambiguous has severely limited the effectiveness of collaborative tagging systems in describing document content and retrieving relevant documents for Web users. For example, when a user wants to retrieve documents about the city of San Francisco from Delicious by using the tag *sf*, documents about science fictions are also returned.

In order to understand the semantics of the tags, we need to know the contexts in which they are used. While a tag itself offers little information on this matter, its associations with other tags, users and documents in a folksonomy provide valuable clues for understanding its semantics. We discover in a preliminary study [2] that users are rather consistent in using a tag to represent a particular concept – if a user uses *sf* to represent ‘San Francisco’, he or she is less likely to use the same tag to represent other concepts such as ‘science fiction’. Based on this observation, we believe that the collective user behaviour observed in a collaborative tagging system serves as an excellent source of information about the semantics of the tags. In this project, we attempt to analyse such user behaviour around ambiguous tags and investigate how clustering algorithms can be employed to disambiguate tags.

2. Background and Motivation

Tagging originates from the idea of using keywords to describe and index resources [12]. Collaborative tagging systems take this idea further by allowing general users to assign freely-chosen tags to Web resources. For example, one can post a bookmark of the homepage of BBC, *http://www.bbc.co.uk/*, to Delicious, and assign to it tags such as *tv*, *media* and *sports*. As tags of different users are aggregated, the tags form a kind of signature of the document which acts as an overall description of the page to be used in future retrieval.

Collaborative tagging systems have started to thrive and grow in number since late 2003 and early 2004 [8]. Col-

[1] http://delicious.com/

[2] http://www.bibsonomy.org/

laborative tagging is generally considered to have a number of advantages over traditional methods of organising information [13, 15] as evidently shown by its popularity among Web users and in different applications. In particular, the flexibility and freedom offered by these systems to Web users are what make them distinguishable from traditional systems which make use of predefined taxonomies. However, collaborative tagging also suffers from several limitations due largely to its unrestricted nature. Since vocabulary is uncontrolled, there is no way to make sure that a tag corresponds to a single well-defined concept.

There are attempts to contextualise tags – putting tags in appropriate contexts so that their meanings can be understood – by performing data mining techniques on folksonomies (e.g. [3, 16]). Some projects also attempt to generate hierarchical structures or ontologies of tags from folksonomies (e.g. [18, 22]). An important finding by Mika [14] is that the associations between tags are best captured when the social context in which these tags are used are considered. While the studies mentioned above focus on discovering significant associations between tags as a means of revealing the semantics of tags, the differences between tag associations in different contexts are seldom considered. For example, most proposed methods are unable to tell the differences between the tags associated with, for example, the tag *sf* when it is used in different contexts such as 'San Francisco' and 'science fiction'.

We believe a better framework for studying tag contextualisation is very much desirable. Firstly, this allows us to have a better understanding of the semantics of tags being used in a folksonomy, and gives us a clearer idea of how the tags are actually used in the system, as opposed to what the tags ought to mean. Secondly, by discovering related tags in different contexts, further methods can then be developed to facilitate browsing or retrieving resources in a collaborative tagging system by performing automatic classification of the resources. For example, by examining the tags associated with a bookmark, we can decide whether it is about 'San Francisco' or 'Science Fiction'.

3. Research Methodology

Our approach of solving the problem of tag ambiguity is by studying the collective user behaviour around the usage of the tags. Applying large scale clustering techniques and statistical analysis on folksonomies probably allows us to discover, for example, that *sf* is highly associated with *california*, *bayarea*, *science* and *fiction*. However, what we attempt to find out is, for example, that *sf* is highly associated with *california* and *bayarea* in one occasion and with *science* and *fiction* in another. We achieve this by putting the tags into the social contexts in which they are used.

Firstly, we adopt the network model of folksonomies commonly found in the literature (e.g. [9, 14]). A folksonomy can be considered as a tripartite graph involving three disjoint sets of entities: users, tags and documents. Users are participants of the collaborative tagging system and they assign tags to documents which can be any kind of Web resources such as bookmarks in Delicious. Formally, a folksonomy is defined as follows.

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 documents, and* $A \subseteq U \times T \times D$ *is a set of annotations.*

A is sometimes referred to as a set of *taggings* which represents the fact that a particular user $u \in U$ has assigned a tag $t \in T$ to a document $d \in D$. As we want to focus on a particular tag $t \in T$ in the folksonomy $\mathbf{F}$ in order to understand its semantics, we can extract a subset B_t of the folksonomy by restricting $\mathbf{F}$ to t: $B_t = (U_t, D_t, E)$, where $U_t = \{u|(u,t,d) \in A\}$ is the set of user who have used t, $D_t = \{d|(u,t,d) \in A\}$ is the set of documents which have been assigned t, and $E = \{(u,d)|(u,t,d) \in A\}$. This is in fact a bipartite graph involving the sets of users and documents which are associated with the tag t, with $U \cup D$ as the set of vertices and E as the set of edges. The graph can be represented as a matrix $\mathbf{A} = \{x_{ij}\}$, where $x_{ij} = 1$ if user u_i has used the tag t on document d_j, or if $(u_i, d_j) \in E$.

Two different one-mode networks, one of documents and another of users, can be generated from this bipartite graph using matrix multiplication: $\mathbf{X} = \mathbf{A}^{\mathrm{T}}\mathbf{A}$ and $\mathbf{Y} = \mathbf{A}\mathbf{A}^{\mathrm{T}}$. This process actually creates a similarity graph of documents (users) by calculating their pairwise similarity based on the users (documents) associated with them. The matrix $\mathbf{X}$ represents a network of documents to which some users have assigned the tag t. An edge in this network is actually weighted by the number of unique users who have assigned the tag to both of the documents on the two ends. On the other hand, $\mathbf{Y}$ represents a derived social network in which two users are connected by an edge if they have both assigned the tag on the same document.

By constructing these networks with respect to the tag t, we are in fact putting the tag into the social contexts in which it is used. If users are consistent in using the tag, or in other words if the users always use a tag to mean the same thing, then we should find different groups of closely connected documents in $\mathbf{X}$, and different groups of closely connected users in $\mathbf{Y}$, which correspond to the different contexts in which t is used. Moreover, graph clustering algorithms can be applied to automatically discover the different groups of documents or tags, generating a set of related tags to represent each of the contexts in which the tag t is used. In order to testify whether our assumption of user consistency in tag usage, we have conducted several preliminary experiments on data collected from Delicious.

4. Preliminary Study

The aim of this preliminary study [2] is to investigate whether our proposal of contextualising tags by analysing collective user behaviour is a valid one. In this study, we examine the networks of users and documents associated with the tag *sf*, and attempt to understand how the different contexts in which the tags are used can be identified. The tag is chosen because it is a popular tag and that it is observed to be used to represent multiple concepts in Delicious.

We construct the networks of documents and users using the method we have described in the previous section. When visualising the network using the Kawada-Kawai algorithm [10], we observe two large groups of nodes which are highly connected within the groups and relatively loosely connected between them. We manually examine in particular the documents and classify them as either related to 'San Francisco' or 'science fiction' based on their content as well as the other tags assigned to them by the users. We discover that each of the groups corresponds to one of the two topics. In other words, the documents can be divided into two groups by considering only the pattern of usage of the tag by the users. This shows that users are mostly consistent when using the tag, otherwise documents about totally different topics will be connected to each other, resulting in a network with no distinguishable boundaries between groups of nodes.

To further investigate the effectiveness of our proposal, we also carry out experiments on several other ambiguous tags, such as *bridge*, *opera* and *wine*, by applying the greedy community-discovery algorithm based on the measure of modularity [4] to the networks obtained. Our experiments reveal that in each of the cases clusters of node can always be found. We try to understand what contexts the different clusters correspond to by extracting the most frequently used tags in the clusters, and they actually provide very clear indications of what the clusters are about. For example, for the tag *opera* we find tags such as *browser*, *software* and *web* in one cluster, and tags such as *music*, *classical* and *art* in another, suggesting that *opera* is used to refer to the Web browser as well as a kind of musical performance. Part of the results of these experiments can be found in [1].

The preliminary study gives satisfactory and encouraging results, suggesting that the social contexts in which the tags are used have an important role in defining tag semantics. The following section describes how we are going to carry out evaluation of larger scale so as to investigate the effectiveness and limitations of the proposed method.

5. Method of Evaluation

While the proposed method gives promising results on several examples of ambiguous tags, we would like to further investigate its usefulness in a systematic way through evaluations of larger scale. In particular, we attempt to answer three research questions: (1) does social context provide better information for understand the semantics of tags than common document clustering techniques; (2) which network models best represent the associations between documents in the social context, and what kind of clustering algorithms is most suitable for the tasks; and (3) what is the most suitable representation of the contexts in which the tags are used.

In order to test the usefulness of the proposed method, we have collected data of more than 50 tags from Delicious, most of which are observed to be used in more than one context. We have also recruited users who are familiar with collaborative tagging systems to manually classify documents into different contexts so that we have references against which the results of an automatic algorithm can be compared. Performance measures such as precision, recall and other adopted measures of accuracy will be employed for evaluation.

To answer the research questions mentioned, we plan to carry out several experiments. Firstly, we will compare the performance of our proposed method with document clustering based wholly on keyword similarity, which will give us an idea of whether the social contexts are more useful in revealing the semantics of tags. This will be extended by using different clustering algorithms on the networks obtained. Analysis will be made to find out which algorithm and which values of parameters best suit the task. Furthermore, we will study how the different contexts revealed in the clustering process can be represented. A simple method would involve extracting the most frequently used tags in a cluster. More sophisticated methods such as identifying pairs of tags which are connected by edges with the largest weights can also be used.

In addition, our work is in principle similar to studies of document clustering. This is a problem quite extensively studied in the literature (e.g. [5, 19]) and is also addressed by commercial systems such as Vivisimo [11].[3] Most existing methods extract keywords from documents and calculate their similarity based on the keywords to obtain a set of clusters. Another area of studies relevant to our work is word sense discrimination, which is a sub-topic of word sense disambiguation [17]. Word sense discrimination attempts to determine if two tokens of a word correspond to the same sense or different senses. However, to the best of our knowledge, social contexts are not considered in dealing with these problems. Although it is very difficult to carry out quantitative comparison due to the lack of a golden standard, we will consider qualitative analysis of the differences between our method and the studies in the above two areas.

[3] The public version of Vivismo's Web search engine, Clusty, can be found at http://clusty.com/.

6. Conclusion

This paper outlines our research project in which we propose to perform tag contextualising using the collective user behaviour in a collaborative tagging system. Our preliminary study shows that the proposed method has the potential to be developed into a systematic way of revealing semantics of tags by putting them in appropriate social contexts. We envisage that the results of this project will open up several new research opportunities such as the application of folksonomies in information retrieval on the Web. For example, one application which we would very much like to look into is how the frequently used tags extracted by our method can be used to classify documents returned by search engines when an ambiguous keyword is used in a query. Finally, we believe this project is significant in enhancing our understanding of the dynamics in collaborative tagging systems and how tags are used by Web users.

References

[1] C. M. Au Yeung, N. Gibbins, and N. Shadbolt. Tag meaning disambiguation through analysis of tripartite structure of folksonomies. In *WI-IAT'07: Proc. of the 2007 IEEE/WIC/ACM International Conferences on Web Intelligence and Intelligent Agent Technology - Workshops, Silicon Valley, CA, USA, November, 2007*, pages 3–6. IEEE Computer Society, 2007.

[2] C. M. Au Yeung, N. Gibbins, and N. Shadbolt. Understanding the semantics of ambiguous tags in folksonomies. In L. C. et al., editor, *Proc. of the First International Workshop on Emergent Semantics and Ontology Evolution, ESOE 2007, co-located with ISWC 2007 + ASWC 2007, Busan, Korea, November 12, 2007*, volume 292 of *CEUR Workshop Proceedings*, pages 108–121. CEUR-WS.org, 2007.

[3] G. Begelman, P. Keller, and F. Smadja. Automated tag clustering: Improving search and exploration in the tag space. In *Proc. of the Collaborative Web Tagging Workshop, WWW2006*, 2006.

[4] A. Clauset, M. E. J. Newman, and C. Moore. Finding community structure in very large networks. *Physical Review E*, 70:066111, 2004.

[5] D. R. Cutting, D. R. Karger, J. O. Pedersen, and J. W. Tukey. Scatter/gather: a cluster-based approach to browsing large document collections. In *SIGIR '92: Proc. of the 15th annual international ACM SIGIR conference on Research and development in information retrieval, Copenhagen, Denmark*, pages 318–329. ACM, 1992.

[6] G. W. Furnas, C. Fake, L. von Ahn, J. Schachter, S. Golder, K. Fox, M. Davis, C. Marlow, and M. Naaman. Why do tagging systems work? In *CHI '06 extended abstracts on Human factors in computing systems*, pages 36–39. ACM, 2006.

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

[8] T. Hammond, T. Hannay, B. Lund, and J. Scott. Social bookmarking tools (i): A general review. *D-Lib Magazine*, 11(4), April 2005.

[9] A. Hotho, R. Jäschke, C. Schmitz, and G. Stumme. Information retrieval in folksonomies: Search and ranking. In Y. Sure and J. Domingue, editors, *The Semantic Web: Research and Applications*, volume 4011 of *LNCS*, pages 411–426. Springer, June 2006.

[10] T. Kamada and S. Kawai. An algorithm for drawing general undirected graphs. *Inf. Process. Lett.*, 31(1):7–15, 1989.

[11] S. Koshman, A. Spink, and B. J. Jansen. Web searching on the vivisimo search engine. *Journal of the American Society for Information Science and Technology*, 57(14):1875–1887, 2006.

[12] F. W. Lancaster. *Indexing and abstracting in theory and practice*. Facet Publishing, London, 2003.

[13] A. Mathes. Folksonomies - cooperative classification and communication through shared metadata. *http://www.adammathes.com/academic/computer-mediated-communication/folksonomies.html*, 2004. Retrieved on 3 September 2008.

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

[15] E. Quintarelli. Folksonomies: power to the people. ISKO Italy-UniMIB meeting, June 2005.

[16] C. Schmitz, A. Hotho, R. Jäschke, and G. Stumme. Mining association rules in folksonomies. In V. Batagelj, H.-H. Bock, A. Ferligoj, and A. Ziberna, editors, *Data Science and Classification: Proc. of the 10th IFCS Conference*, Studies in Classification, Data Analysis, and Knowledge Organization, pages 261–270. Springer, July 2006.

[17] H. Schütze. Automatic word sense discrimination. *Comput. Linguist.*, 24(1):97–123, 1998.

[18] L. Specia and E. Motta. Integrating folksonomies with the semantic web. In E. Franconi, M. Kifer, and W. May, editors, *ESWC2007: Proc. of the European Semantic Web Conference, Innsbruck, Austria, June 3-7, 2007*, volume 4519 of *LNCS*. Springer-Verlag, 2007.

[19] J. Stefanowski and D. Weiss. Carrot^2 and language properties in web search results clustering. In E. M. Ruiz, J. Segovia, and P. S. Szczepaniak, editors, *Proc. of First International Atlantic Web Intelligence Conference, AWIC 2003, Madrid, Spain, May 5-6, 2003*, volume 2663 of *LNCS*, pages 240–249. Springer, 2003.

[20] T. V. Wal. Folksonomy definition and wikipedia. *http://www.vanderwal.net/random/entrysel.php?blog=1750*, November 2, 2005. Retrieved on 13 Feb 2008.

[21] X. Wu, L. Zhang, and Y. Yu. Exploring social annotations for the semantic web. In *WWW '06: Proc. of the 15th international conference on World Wide Web, Edinburgh, Scotland, May 23-26, 2006*, pages 417–426. ACM Press, 2006.

[22] M. Zhou, S. Bao, X. Wu, and Y. Yu. An unsupervised model for exploring hierarchical semantics from social annotations. In K. A. et al., editor, *The Semantic Web, 6th International Semantic Web Conference, 2nd Asian Semantic Web Conference, ISWC 2007 + ASWC 2007, Busan, Korea, November 11-15, 2007*, volume 4825 of *LNCS*, pages 680–693. Springer, 2007.