Tag Meaning Disambiguation through Analysis of Tripartite Structure of Folksonomies

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

Abstract

Collaborative tagging systems are becoming very popular recently. Web users use freely-chosen tags to describe shared resources, resulting in a folksonomy. One problem of folksonomies is that tags which appear in the same form may carry multiple meanings and represent different concepts. As this kind of tags are ambiguous, the precisions in both description and retrieval of the shared resources are reduced. We attempt to develop effective methods to disambiguate tags by studying the tripartite structure of folksonomies. This paper describes the network analysis techniques that we employ to discover clusters of nodes in networks and the algorithm for tag disambiguation. Experiments show that the method is very effective in performing the task.

1 Introduction

In collaborative tagging systems [4], users are allowed to choose any keyword they like as tags to describe shared resources. As more tags are aggregated, a kind of classification scheme called a folksonomy [9] starts to take shape. One problem in these systems is that tags which appear in the same form may carry multiple meanings and may be used to represent different concepts. As this kind of tags are ambiguous in meaning, the precisions in both description and retrieval of the shared resources are reduced.

In this paper, we describe our attempt to develop an effective method to disambiguate tags by studying the tripartite structure of folksonomies [5]. We present the network analysis techniques that we employ to discover clusters of nodes in networks. Based on the idea that documents corresponding to the same meaning of a tag would tend to be clustered together, we implement an algorithm to disambiguate tags and carry out experiments to evaluate its effectiveness with data obtained from the social bookmarking site del.icio.us [1].

2 Motivations

Although the freedom offered by collaborative tagging systems to use any tags to describe resources has contributed to its success, such unorganized use of tags has also resulted in a number of problems. In particular, ambiguous tags are quite abundant in folksonomies. For example, in del.icio.us *sf* is used to refer to both the city of San Francisco and to science fictions. The ambiguity of tags poses challenges to applications such as retrieval of relevant resources and matching of user interests. While quite a number of authors have worked on the problem of discovering synonymous tags (e.g. [7]), few can be found to address the problem of tag ambiguity. Wu et al. [10] describe an algorithm to discover the different dimensions of knowledge existing in a folksonomy. They employ statistical analysis on folksonomies, and study the conditional probabilities of tags in different conceptual dimensions. Tags with multiple meanings will then score high in more than one dimensions in the conceptual space. However, one limitation of their method is that the number of dimensions must be determined beforehand. The facts that tag ambiguity is a major problem that affects the effectiveness and efficiency of collaborative tagging systems, and that few authors have addressed the problem, give us the motivation to develop useful methods to tackle the problem of tag ambiguity.

3 Tripartite Structure of Folksonomies

A folksonomy is generally agreed to be consisting of at least three sets of elements [5, 10], namely users, tags,

[1] http://del.icio.us/

Tag	Triples	Documents	Users
sf	238,117	427	19,979
opera	148,484	313	25,907
cambridge	61,424	242	13,455
tube	382,826	502	74,527

Table 1. Data Collected for Experiments on Tag Meaning Disambiguation.

and resources. Since we are focusing on tagging data in del.icio.us, resources are primarily Web documents. In this paper, we adopt the definition of folksonomy proposed by Mika [5].

Definition 1 *A folksonomy F is tuple $F = (U, T, D, A)$, 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.*

As there are three sets of elements in a folksonomy, a tripartite structure can be constructed based on the associations between these elements. However, by focusing on one of these three elements, we are able to fold the tripartite structure into a bipartite one [5], which allows us to perform analysis more easily. Since we are dealing with tags and their meanings in this paper, we will concentrate on the bipartite graphs obtained by focusing on tags.

By focusing on a single tag, we obtain a bipartite graph UD_t with respect to a particular tag t:

$$UD_t = \langle U \cup D, E_{ud} \rangle, E_{ud} = \{(u, d) | (u, t, d) \in A\}$$

An edge exists between a user and a document if the user has assigned the tag t to the document. The graph can be represented in matrix form, which we denote as $\mathbf{Y} = \{y_{ij}\}$, $y_{ij} = 1$ if there is an edge connecting u_i and d_j, and $y_{ik} = 0$ otherwise. This bipartite graph can be folded into two one-mode networks, which we denote as $\mathbf{S} = \mathbf{Y}\mathbf{Y}'$, and $\mathbf{C} = \mathbf{Y}'\mathbf{Y}$. The matrix $\mathbf{S}$ shows the affiliation between the users who have used the tag t, weighted by the number of documents to which they have both assigned the tag. Users who use the tag for the same meaning are likely to be connected with each other. On the other hand, $\mathbf{C}$, with the edges weighted by the number of users who have assigned tag t to both documents, is likely to connect documents which are related to the same sense of the given tag.

4 Tag Meaning Disambiguation

Although tags may carry different meanings, one can still single out the particular meanings when we examine the Web documents to which the tags are assigned. The documents as well as the other tags associated with them provide the context to understand an ambiguous tag. We observe that Web documents which correspond to the same meaning of a tag tend to be grouped together to form clusters. By revealing these clusters and examining the documents and tags involved, it is possible that the different meanings of a tag can be discovered.

4.1 Discovery Community Structures in Networks

A cluster in the network is basically a group of nodes in which nodes have denser connections with each other than with nodes in other clusters. Such task is usually referred to as the problem of discovering community structures within networks [3]. Recently Girvan and Newman [6] introduce a new algorithm, now generally referred to as the GN algorithm, to tackle the problem. The algorithm is a divisive one as it attempts to remove edges in the network in a progressive manner until the underlying community structure is revealed. The decision to remove an edge is made based on the value of its "edge betweenness," which is defined as the number of shortest paths between pairs of nodes that run along it. The authors further propose the notion of modularity as a measure of the goodness of a particular division of a network [6]. Thus, the aim of the algorithm becomes maximizing the value of modularity. The GN algorithm has been demonstrated to be highly effective on both artificially generated and real world networks. It has also been widely adopted in recent years because it overcomes many of the shortcomings of traditional methods [8]. Hence, we try to apply it to our problem of tag disambiguation.

4.2 Proposed Method

We develop our method for tag meaning disambiguation based on the GN algorithm and the notion of modularity. It involves the following steps

1. Collect tagging data that involves *t* and construct a one-mode network of documents out of the tagging data.
2. Calculate edge betweenness and remove the edge with the highest value.
3. Calculate the modularity of the current division of the network and update the best division and the highest value of modularity obtained so far.
4. Repeat Steps 2 to 3 until no more edges remain in the network. The division with the highest value of modularity is obtained.
5. For each of the clusters in the final division of the network, obtained the 10 most frequently used tags among the documents. This set of tags serve as a signature of the cluster.

Although the algorithm we described above would not produce exactly the different meanings of a tag, the most frequently used tags in a cluster should provide a coherent context from which the exact meaning of the tag can be easily deduced. In the following section, we will apply this algorithm to several ambiguous tags in del.icio.us and evaluate its effectiveness in tag meaning disambiguation.

5 Experiments

To evaluate the algorithm we described above, we carry out experiments on four tags in del.icio.us which are observed to have multiple meanings or usages. The four tags are *sf*, *opera*, *cambridge* and *tube*. We collect tagging data from the del.icio.us website by using a crawler program. The data includes documents which are tagged by the chosen tags, the users who have tagged these documents, and the other tags that have been assigned to them. Table 1 summarizes the statistics of the data.

We apply the algorithm on the network of documents, and obtain the top ten tags of different clusters of documents. We feed the results into Pajek [2] by which we obtain the visualizations of the networks after the clustering process. The results are shown in Table 2. The first column in the tables refers to cluster numbers. The second column lists the 10 most frequently used tags in the clusters. The visualizations of the networks are shown in Fig. 1. The numbers labelling the nodes in the networks indicate the cluster to which the nodes belong.

In the tables and the figures, we can observe that different clusters of documents correspond to different meanings of the tags. For the tag *sf*, we identify the two different meanings of the tag, namely "San Francisco" and "science fiction." For the tag *opera*, we discover its meaning in both the context of the Web and the context of musical performance.[2] Experiment on the tag *cambridge* shows that it is used to refer to the area near Boston in the United States, the city in England and the university in the city. Finally, the experiment on the tag *tube* reveal several meanings of the tag, including the underground rail network in London (Cluster 1), a kind of electronic components (Cluster 2 and 7), and also the video sharing site Youtube (Cluster 3 to 6).

The experiments show that our proposed method can actually be used to find out the different meanings of an ambiguous tag. It is clear that the method is able to give us a better understanding of the ambiguous tags in folksonomies and how these tags are used by different users in the system. This should also be beneficial to further applications which aim at extracting semantics from folksonomies to be used in the Semantic Web. [1].

[2]Opera is the name of a Web browser: http://www.opera.com/, while opera is also a form of musical performance.

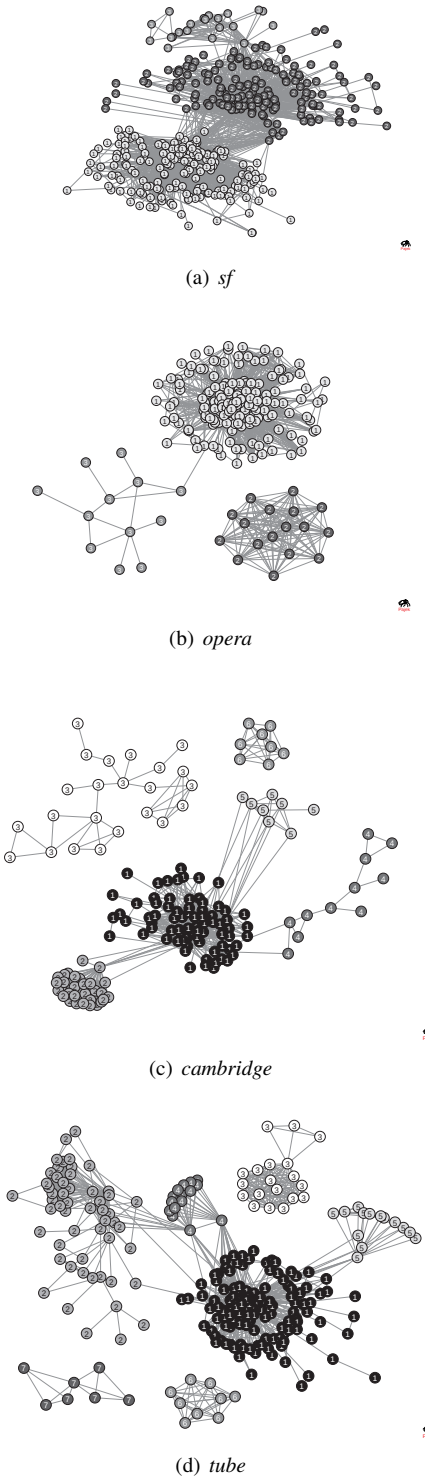
Tag: *sf*	
Cluster	**10 Most Frequently Used Tags**
1	sf, scifi, fiction, books, sci-fi, writing, literature, science, sciencefiction, fantasy
2	sf, sanfrancisco, bayarea, san, francisco, california, travel, events, art, san_francisco
3	sf, sanfrancisco, design, bayarea, blog, food, todo, california, shopping, san

Tag: *opera*	
Cluster	**10 Most Frequently Used Tags**
1	opera, browser, web, software, javascript, browsers, tips, tools, internet, firefox
2	opera, shopping, imported, shop, design, store, home, inspiration, work, personal
3	opera, music, musique, classical, art, culture, musica, música, classic, travel

Tag: *cambridge*	
Cluster	**10 Most Frequently Used Tags**
1	cambridge, university, uk, england, science, cam, local, cambridgeuniversity, research, community
2	cambridge, bcc_school, activism, education, community, contact, bcc, politics, critical_economy, blog
3	cambridge, boston, restaurants, food, massachusetts, imported, local, restaurant, venues, clubs
4	cambridge, english, cpe, cae, boston, online, fce, exam, inglés, esl
5	cambridge, mappingurbanism, visualisation, design, social, information, maps, mapping, infovis, toread
6	cambridge, letting, uk, photography, search, property, flats, cambsproperty, financial, fundraising

Tag: *tube*	
Cluster	**10 Most Frequently Used Tags**
1	tube, london, underground, travel, transport, maps, uk, map, subway, reference
2	tube, diy, audio, electronics, amp, amplifier, amps, tubes, guitar, music
3	tube, video, web, internet, tv, online, web2.0, media, videos, imported
4	tube, video, youtube, videos, funny, cool, interesting, sport, fun, humor
5	tube, video, videos, online, web2.0, youtube, free, media, movie, fun
6	tube, youtube, video, videos, cool, feel.good, fun, funny, flash, music
7	tube, radio, electronics, tubes, antique, amplifier, data, audio, info, incarnate

Table 2. Results of the experiments on tag meaning disambiguation.



(a) *sf*

(b) *opera*

(c) *cambridge*

(d) *tube*

Figure 1. Result of clustering in the networks of documents.

6 Conclusion

In this paper we present a method to discover the different meanings of an ambiguous tags in a folksonomy. We preform experiments on four ambiguous tags, and the results show that the method is very effective in tag disambiguation. However, some issues, such as that a particular meaning of a tag can be observed in more than one cluster, remain to be investigated. In the future, we will carry out more analysis to study the effectiveness of the algorithm and study how the result of the algorithm can be refined to be used in automatic tag disambiguation.

References

[1] T. Berners-Lee, J. Hendler, and O. Lassila. The semantic web. *Sci. Am.*, 284(5):34–43, 2001.

[2] Wouter de Nooy, Andrej Mrvar, and Vladimir Batagelj. *Exploratory Social Network Analysis with Pajek (Structural Analysis in the Social Sciences)*. Cambridge University Press, January 2005.

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

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

[5] Peter Mika. Ontologies are us: A unified model of social networks and semantics. In *International Semantic Web Conference*, pages 522–536, 2005.

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

[7] Satoshi Niwa, Takuo Doi, and Shinichi Honiden. Folksonomy tag organization method based on the tripartite graph analysis. In *IJCAI Workshop on Semantic Web for Collaborative Knowledge Acquisition*, 2007.

[8] 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.

[9] G. Smith. Atomiq: Folksonomy: Social classification. *http://atomiq.org/archives/2004/08/folksonomy_social_classification.html*, 2004.

[10] Xian Wu, Lei Zhang, and Yong Yu. Exploring social annotations for the semantic web. In *WWW '06: Proceedings of the 15th international conference on World Wide Web*, pages 417–426, New York, NY, USA, 2006. ACM Press.