

# Understanding the Semantics of Ambiguous Tags in Folksonomies

Ching-man Au Yeung, Nicholas Gibbins, and Nigel Shadbolt

Intelligence, Agents and Multimedia Group (IAM),  
School of Electronics and Computer Science,  
University of Southampton,  
Southampton SO17 1BJ, UK  
{cmay06r,nmg,nrs}@ecs.soton.ac.uk

**Abstract.** The use of tags to describe Web resources in a collaborative manner has experienced rising popularity among Web users in recent years. The product of such activity is given the name folksonomy, which can be considered as a scheme of organizing information in the users' own way. In this paper, we present a possible way to analyze the tripartite graphs – graphs involving users, tags and resources – of folksonomies and discuss how these elements acquire their meanings through their associations with other elements, a process we call mutual contextualization. In particular, we demonstrate how different meanings of ambiguous tags can be discovered through such analysis of the tripartite graph by studying the tag *sf*. We also discuss how the result can be used as a basis to better understand the nature of folksonomies.

## 1 Introduction

The use of freely-chosen words or phrases called tags to classify Web resources has experienced rising popularity among Web users in recent years. Through the use of tags, Web users come to share and organize their favourite Web resources in different social tagging systems, such as del.icio.us<sup>1</sup> and Flickr<sup>2</sup>. The result of this collaborative and social tagging activity is given the name folksonomy, which refers to the classification system evolved from the individual contributions of tags from the users [1].

Collaborative tagging possesses a number of advantages which account for its popularity. These include its simplicity as well as the freedom enjoyed by the users to choose their own tags. However, some limitations and shortcomings, such as the problem of ambiguous meanings of tags and the existence of synonyms, also affect its effectiveness to organize resources on the Web. As collaborative tagging attracts the attentions of researchers, methods on how useful information can be discovered from the seemingly chaotic folksonomies have been developed. In particular, some focus on discovering similar documents or communities of

<sup>1</sup> <http://del.icio.us/>

<sup>2</sup> <http://www.flickr.com/>

shared interests [17, 13], while some perform analysis on the affiliation between entities to find out different relations between tags [10, 14].

In this paper we focus on analysis of tripartite graphs of folksonomies, graphs which involve the three basic elements of collaborative tagging, namely users, tags and resources. We present how these elements come to acquire their own semantics through their connections with other elements in the graphs, a process which we call mutual contextualization. In particular, we carry out a preliminary study on tripartite graphs with data obtained from del.icio.us, and demonstrate how we can understand the semantics of ambiguous tags by examining the structures of these graphs. We also discuss how the result can be used as a basis to acquire a better understanding of the nature of folksonomies.

The rest of this paper is structured as follows. Section 2 gives some background information on collaborative tagging systems and folksonomies. We describe the process of mutual contextualization between the three basic elements in Section 3. We detail the preliminary study on tripartite graphs of folksonomies in Section 4, followed by discussions in Section 5. Finally we present our conclusions and discuss possible future research directions in Section 6.

## 2 Background

### 2.1 Collaborative Tagging Systems

Tagging originates from the idea of using keywords to describe and classify resources. These keywords are descriptive terms which indicate the topics addressed by the resources. Collaborative tagging systems emerged in recent years have taken this idea further by allowing general users to assign tags, which are freely-chosen keywords, to resources on the Web. For example, one can store a bookmark of the page “<http://www.google.com/>” on a collaborative tagging system, and assign to it the tags *google*, *search* and *useful*. As the tags of different users are aggregated, the tags form a kind of signature of the document, which can be used for future retrieval or indication of the nature of the page.

Collaborative tagging systems have started to thrive and grow in number since late 2003 and early 2004 [6]. As one of the earliest initiative of collaborative tagging, del.icio.us provides a kind of social bookmarking service, which allows users to store their bookmarks on the Web, and use tags to describe them. Other services focusing on different forms of Web resources appeared shortly. For example, Flickr allows users to tag digital photos uploaded by themselves.

Collaborative tagging are generally considered to have a number of advantages over traditional methods of organizing information, as evidently shown by its popularity among general Web users and its application on a wide range of Web resources. The following features of collaborative tagging are generally attributed to their success and popularity [1, 15, 18].

*Low cognitive cost and entry barriers* The simplicity of tagging allows any Web user to classify their favourite Web resources by using keywords that are not constrained by predefined vocabularies.

*Immediate feedback and communication* Tag suggestions in collaborative tagging systems provide mechanisms for users to communicate implicitly with each other through tag suggestions to describe resources on the Web.

*Quick Adaptation to Changes in Vocabulary* The freedom provided by tagging allows fast response to changes in the use of language and the emergency of new words. Terms like *AJAX*, *Web2.0*, *ontologies* and *social network* can be used readily by the users without the need to modify any pre-defined schemes.

*Individual needs and formation of organization* Tagging systems provide a convenient means for Web users to organize their favourite Web resources. Besides, as the systems develop, users are able to discover other people who are also interested in similar items.

On the other hand, limitations and problems of existing collaborative tagging systems have also been identified [1, 13, 18]. These issues hinder the growth or affect the usefulness of the systems.

*Tag Ambiguity* Since vocabulary is uncontrolled in collaborative tagging systems, there is no way to make sure that a tag is corresponding to a single and well-defined concept. For an example, items being tagged by the term *sf* may either be related to something about science fiction or the city San Francisco.

*The use of multiple words and spaces* Some systems allow users to input tags separated by spaces. Problems arise when users would like to use phrases with multiple words to describe the Web resources.

*The problem of synonyms* Different tags can be used to refer to the same concept in a tagging system. For example, “mac,” “macintosh,” and “apple” can all be used to describe Web resources related to Apple Macintosh computers[1]. The use of different word forms such as plurals and parts of speech also exacerbate the problem.

*Lack of semantics* A tag provides limited information about the documents being tagged. For example, when tagging an URL with the tag “podcast,” one can mean that the website provides podcast, describes the use of podcast, or provides details on the history of podcasting.

## **2.2 Folksonomies**

As more tags are contributed to a collaborative tagging system by the users, a form of classification scheme will take shape. Such scheme emerges from the collective efforts of the participating users, reflecting their own viewpoints on how the shared resources on the Web should be described using various tags. This product of collaborative tagging is now commonly referred to as *folksonomy* [16]. A folksonomy is generally agreed to be consisting of at least the following three sets of entities [9, 10, 18].

*Users* Users are the ones who assign tags to Web resources in social tagging systems. They are also referred to as actors, as in social network analysis.

*Tags* Tags are keywords chosen by users to describe and categorize resources. Depending on systems, tags can be a single word, a phrase or a combination of symbols and alphabets. Tags are referred to as concepts in some works [10].

*Resources* Resources refer to the objects that are being tagged by the users in the social tagging systems. Depending on the system, resources can be used to refer to Web pages (bookmarks) as in del.icio.us or photos as in Flickr. Resources are also referred to as instances, objects or documents, depending on the context.

Quite a number of research works perform analysis on social tagging systems. However, even though most works adopt a model involving the above three entities, with a few mentioning extra dimensions such as the time of tagging, there is actually not a common consensus on the formal definition of folksonomy. Below we summarize the attempts in this respect.

Mika [10] represents a social tagging system as a tripartite graph, in which the set of vertices can be partitioned into three disjoint sets  $A$ ,  $C$  and  $I$ , corresponding to the set of actors, the set of concepts and the set of objects being tagged. A folksonomy is then defined by a set of annotations  $T \subseteq A \times C \times I$ , an element of which is a triple representing an actor assigning a concept to an object being tagged.

Gruber [5] proposes a “tag ontology” which formalizes the activity of tagging through the use of an ontology. He suggests that tagging can be defined using a five-place relation:  $Tagging(object, tag, tagger, source, [+/-])$ , with object being the Web resources being tagged, tagger being the user who assigns tags, source being the system from which this annotation originates, and  $[+/-]$  representing either a positive or negative vote placed on this annotation by the tagger. Newman [12] also developed a similar ontology for tagging. The act of tagging is modelled as a relation  $T(Resource, Tagging(Tag, Agent, Time))$ .

Hotho et al. [7] define a folksonomy as a tuple  $\mathbb{F} := (U, T, R, Y, \prec)$ . The finite sets  $U$ ,  $T$  and  $R$  correspond to the set of users, tags and resources respectively.  $Y$  refers to the tag assignments, which are ternary relation between the above three sets:  $Y \subseteq U \times T \times R$ .  $\prec$  is a user-specific relation which defines the sub/superordinate relations between tags. By dropping  $\prec$ , the folksonomy can be reduced to a tripartite graph, which is equivalent to Mika’s model.

### 3 Mutual Contextualization in Folksonomies

The power of folksonomies lies in the interrelations between the three elements. A tag is only a symbol if it is not assigned to some Web resources. A tag is also ambiguous without a user’s own interpretation of its meaning. Similarly, a user, though identified by its username, is characterized by the tags it uses and the resources it tags. Finally, a document is given semantics because tags act as a

form of metadata annotation. Hence, it is obvious that each of these elements in a folksonomy would be meaningless, or at least ambiguous in meaning, if they are considered independently. In other words, the semantics of one element depends on the context given by the other two, or all, elements that are related to it.

To further understand this kind of mutual contextualization, we examine each of the three elements in a folksonomy in detail. For more specific discussions, we assume that the Web resources involved are all Web documents. In addition, we define the data in a social tagging system, a folksonomy, as follows.

**Definition 1.** *A folksonomy  $F$  is a 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.*

By adopting this definition, we are actually using the model described by Mika [10]. Since we are mainly focusing on the associations between the three elements and are obtaining data from a single social bookmarking site, information such as the time stamps and sources of tagging is irrelevant here. Thus, the definition we used here is a simple but sufficient one for our work presented here.

As we have mentioned, the three elements forming the tripartite graph of a social tagging system are users, tags and documents (resources). The tripartite graph can be reduced into a bipartite graph if, for example, we focus on a particular tag and extract only the users and documents associated with it. Since there are three types of elements, there can be three different types of bipartite graphs. This step is similar to the method introduced by Mika [10]. However, we distinguish our method from that presented by Mika by focusing on only one instance of a type (e.g. tags), instead of all the items of the same type, allowing us to acquire more specific understanding of the semantics of the instance.

### 3.1 Users

By focusing on a single user  $u$ , we obtain a bipartite graph  $TD_u$  defined as follows:

$$TD_u = \langle T \cup D, E_{td} \rangle, E_{td} = \{\{t, d\} | (u, t, d) \in A\}$$

In other words, an edge exists between a tag and a document if the user has assigned the tag 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  $d_j$ . The bipartite graph represented by the matrix can be folded into two one-mode networks [10]. We denote one of them as  $\mathbf{P} = \mathbf{X}\mathbf{X}'$ , and another as  $\mathbf{R} = \mathbf{X}'\mathbf{X}$ .

$\mathbf{P}$  represents a kind of semantic network which shows the associations between different tags. It should be note that this is unlike the lightweight ontology mentioned in [10], as it only involves tags used by a single user. In other words, this is the personal vocabulary, a personomy [7], of a particular user.

The matrix  $\mathbf{R}$  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 having higher weights on the links between them are those that are considered by the particular user as more related.

### 3.2 Tags

By using a similar method as described above, we can obtain a bipartite graph  $UD_t$  regarding to a particular tag  $t$ :

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

In words, an edge exists between a user and a document if the user has assigned the tag  $t$  to the document. The graph can once again 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  $d_j$ . 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. Since a tag can be used to represent different concepts (such as *sf* for *San Francisco* or *Science Fiction*), and a document provides the necessary content to identify the contextual meaning of the tag, this network is likely to connect users who use the tag for the same meaning.

$\mathbf{C}$  can be considered as another angle of viewing the issue of polysemous or homonymous tags. Thus, with the edges weighted by the number of users who have assigned tag  $t$  to both documents, this network is likely to connect documents which are related to the same sense of the given tag.

### 3.3 Documents

Finally, a bipartite graph  $UT_d$  can also be obtained by considering a particular document  $d$ . The graph is defined as follows:

$$UT_d = \langle U \cup T, E_{ut} \rangle, E_{ut} = \{\{u, t\} | (u, t, d) \in A\}$$

In words, an edge exists between a user and a tag if the user has assigned the tag to the document  $d$ . The graph can be represented in matrix form, which we denote as  $\mathbf{Z} = \{z_{ij}\}$ ,  $z_{ij} = 1$  if there is an edge connecting  $u_i$  and  $t_j$ . Like in the cases of a single user and a single tag, this bipartite graph can be folded into two one-mode networks, which we denote as  $\mathbf{M} = \mathbf{Z}\mathbf{Z}'$ , and  $\mathbf{V} = \mathbf{Z}'\mathbf{Z}$ .

The matrix  $\mathbf{M}$  represent a network in which users are connected based on the documents commonly tagged by them. Since a document may provide more than one kind of information, and users do not interpret the content from a single perspective, the tags assigned by different users will be different, although tags related to the main theme of the document are likely to be used by most users. Hence, users linked to each other by edges of higher weights in this network are more likely to share a common perspective, or are more likely to concern a particular piece of information provided by the document.

On the other hand, the matrix  $\mathbf{V}$  represents a network in which tags are connected and weighted by the number of users who have assigned them to the document. Hence, the network is likely to reveal the different perspective of the users from which they interpret the content of the document.

We can see that different relations between the users, the tags and the documents in a folksonomy will affect how a single user, tag or document is interpreted in the system. Each of these elements provide an appropriate context such that the semantics of the elements can be understood without ambiguity.

## 4 Semantics of Ambiguous Tags

One problem in the existing collaborative tagging system is the existence of ambiguous tags. By “ambiguous tags,” we refer to tags that are intended to represent different concepts by the users. For example, in del.icio.us the tag *sf* has been used to describe documents which are related to science fiction and San Francisco. Another example is the tag *opera*, which are used for describing contents related to opera as a kind of musical performance as well as those related to the WWW browser which is named “Opera.”<sup>3</sup>

As we have discussed, the semantics of a tag depends on the context given by the users who have used it as well as the documents being tagged. By studying the associations between the tag, the users and the documents, we may determine the different meanings of a tag by placing it in the right context. As an illustrative example, we present an analysis of the bipartite graphs obtained from a single tag, which we have chosen for its common occurrence and multiple equally-frequent meanings in order to preserve the clarity of the example. In particular, we would like to find out if it is possible to disambiguate a tag by studying its association with different users and documents.

### 4.1 Understanding a Single Tag

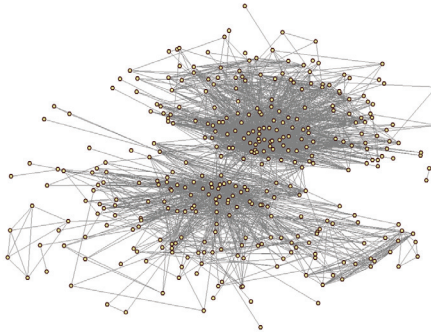
In the experiment described below, we try to examine the networks of users and documents associated with the tag *sf*, and attempt to understand how different interpretation of the tag can be discovered from the analysis of the networks.

The reasons of choosing the tag *sf* as an illustrating example are twofold. Firstly, *sf* is a tag used very frequently by users in del.icio.us. Although the exact number of times that the tag has been used cannot be known from the system, we are able to collect over 5000 triples which involves the tag *sf*. Secondly, by observation, the tag *sf* has been used by users to refer to two very distinctive concepts, namely “science fiction” and “San Francisco.” We expect that users using the tag to refer to one of the two concepts do not use it to refer to the other one. Hence, the tag *sf* is more worthwhile to be examined, and we expect that experiments on the tag can produce clearer results for performing analysis.

In March 2007, data was collected from the del.icio.us website by using a crawler program written in Python. The program retrieved pages listing all bookmarks that have been tagged with *sf*, and subsequently retrieved the published RSS file of each bookmark to obtain the corresponding users and tags associated with it. In other words, the crawler retrieved bookmarks in del.icio.us which have

---

<sup>3</sup> <http://www.opera.com/>



**Fig. 1.** A network of documents tagged by *sf*.

been tagged with *sf*, along with the users who tagged the page, and the tags, including *sf*, they used. In total, 238,117 triples were obtained, each involving a user, an URL of the bookmark, and a tag. A total of 427 distinctive URLs and 19979 users are involved. Out of these triples, 5852 involves the tag *sf*.

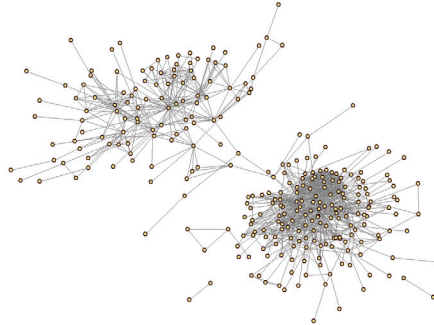
We extract all those triples that involve the tag *sf*, and construct the matrix  $\mathbf{Y}$ , representing the associations between users and bookmarks (documents). We then construct the matrices  $\mathbf{S} = \mathbf{Y}\mathbf{Y}'$ , corresponding to the network of users, and  $\mathbf{C} = \mathbf{Y}'\mathbf{Y}$ , corresponding to the network of documents.

The matrices  $\mathbf{S}$  and  $\mathbf{C}$  are fed into the network analysis package Pajek [3], and visualized as networks. Since some users do not have any associations with other users, as in the case of documents, isolated nodes are removed from the networks. The results are shown in Fig 1 and Fig 2. In Fig 1, nodes represent documents, and two nodes are connected by an edge if a user has tagged both documents with the tag *sf*. Edges are weighted by the number of such users, and is not shown in the figure. In Fig 2, nodes represent users, and two nodes are connected by an edge if both users have tagged a document with the tag *sf*. Edges are weighted by the number of such documents. The networks are visualized using the Kamada-Kawai layout algorithm [8] implemented in Pajek.

Two large clusters of nodes can be observed in both of the networks in Fig 1 and Fig 2. However, as shown in the two figures, there are more connections between the two clusters in the network of documents than in that of users. One hypothesis that can be used to explain the existence of clusters in the network of documents is that they correspond to groups of documents related to the different senses of the tag *sf*. A similar hypothesis that can be applied to the network of users is that the different clusters corresponds to groups of users who have used the tag *sf* to represent different concepts.

Since documents are connected if a user tagged them with the tag *sf*, it implies that connected documents are considered by the user as all related to certain concept represented by the tag *sf*. In addition, if we assume that a user





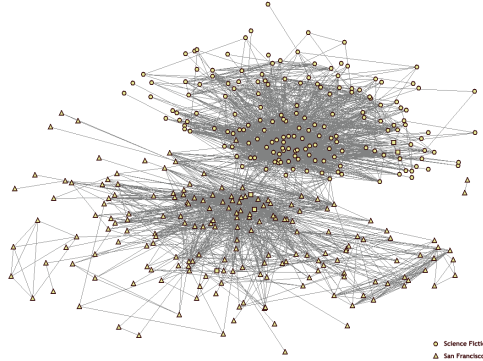
**Fig. 2.** The network of users who used the tag *sf*.

would be consistent in using the same tag for the same concept, it is reasonable to suggest that documents in different clusters would address a different concept represented by the tag *sf*. As we understand through observation that two major concepts – “science fiction” and “San Francisco” are associated with the tag *sf*, we can further suggest that the two major clusters in the network correspond to documents on science fiction and San Francisco respectively. To testify this hypothesis, we perform further analysis on the tagging data.

Firstly, we manually examine all the 357 websites represented by the nodes in the network of documents. We classify the websites into either related to science fiction or San Francisco, based on the content of the website as well as other tags used by the users. We indicate that the website cannot be classified into either of these categories if not enough information or evidence is available. After that, we combine the information with the original network, and use Pajek to draw a new network, as shown in Fig 3.

In the figure, circular nodes represent documents related to science fictions, and triangular nodes represent documents related to San Francisco. Documents that cannot be classified are represented by rectangular nodes. We can see that these two types of nodes are clearly grouped into two clusters. The result shows that the two clusters indeed correspond to two sets of documents related to two distinctive meaning of the tag *sf*.

However, it is interesting to note that there are actually a lot of edges connecting nodes from different clusters. Since nodes are connected if a user tagged them with the tag *sf*, these connections imply that some users actually used the same tag to represent two distinctive concepts. This also explains why the two clusters in the network of users are connected by a few edges. The documents connected by edges between clusters in the network of documents are then responsible for the edges connecting the users from different clusters in the network of users. However, since it would be very difficult to judge accurately whether

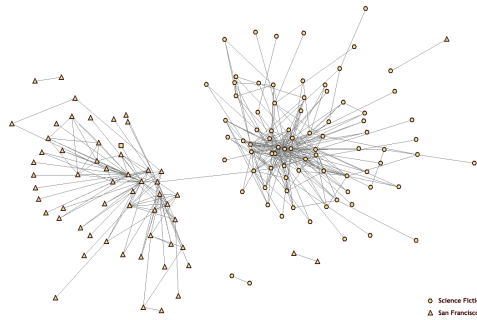


**Fig. 3.** The network of documents tagged by *sf* with classified nodes.

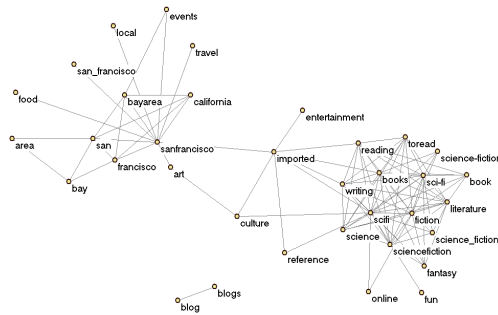
a user always uses the tag *sf* to refer to science fictions or San Francisco, we refrain from performing a similar classification of the users.

To further investigate whether there are many users who actually used the tag to refer to more than one concept, we construct one more network of documents. Based on the data which generates Fig 3, we remove edges which has a weight less than 2. By doing that we effectively ignore all the edges which correspond to cases in which only one user has used the tag *sf* on both of the documents connected by an edge. We also remove nodes that are not connected to any other nodes afterwards. The result is shown in Fig 4. it can be seen that there remains only one edge which connects nodes across the two clusters.

Finally, we examine how different tags are associated with each other given this set of documents and users. Since the documents are all tagged by the tag *sf*, all the other tags can be considered to be related to it. Given the two distinctive concepts represented by the tag, it is reasonable to hypothesize that the tags related to it can also be divided into two groups, one being related to science fictions, and another to San Francisco. We construct a matrix  $\mathbf{T} = \{t_{ij}\}$  to represent the associations between the tags.  $t_{ij}$  is the number of times  $tag_i$  and  $tag_j$  have been used on the same document. Since there are over 8000 unique tags in the data, and many of them have been only used on a few documents, we only concentrate on 35 tags which are used most frequently along with *sf*. The associations between the tags are visualized in Fig 5. We can see that tags which are related to San Francisco are grouped in one cluster while tags related to science fictions are grouped in another cluster. This suggests that we can examine the related tags in order to obtain the different meanings of an ambiguous tag.



**Fig. 4.** The network of documents tagged by *sf* after removal of edges with weights less than two and unconnected nodes.



**Fig. 5.** The network of 35 tags which are most frequently used along with *sf*.

## 5 Discussions

The experiment results show that by analyzing the tripartite graph of folksonomy and the relations between tags, users and documents, we can discover how tags are being used, and better understand the meanings of the tags which are used for multiple meanings. Hence, although the same tag can be used to represent different concepts, the documents and the users still provide the context for understanding specific meanings of the tag. Given the above results, we come to understand more about the characteristics of folksonomies.

### 5.1 Ambiguous Tags from Users' Point of View

Based on the facts that documents of similar topics are clustered together, and that documents are connected by users who have applied the tag *sf*, we see that

the majority of users use the tag to refer to one concept only. This is because if users use the tag arbitrarily to refer to any of the two concepts, we would not be able to observe two clusters in the network. Hence, although a tag can possess several distinctive meanings, users tend to be consistent in referring to the same meaning when they use the tag. One may also suggest that users interested in one concept represented by the tag are not interested in the other, thus producing the two clusters of documents. However, given that the different senses of the tags we examined do not actually have conflicts with each other, and that the experiments actually involves quite a large number of users, it is more reasonable to suggest that consistence in usage is the reason of the clear distinction that we have observed. Hence, this shows that it is possible to understand whether a tag has multiple senses by examining the associations between users and documents.

## 5.2 Existence of Sub-communities

In the experiment, in addition to the two large clusters of nodes, we can also observe within the clusters that there are some nodes which tend to be grouped with each other to form smaller clusters. For example, in Fig 3 on the left and right ends of the clusters of triangular nodes, we can observe that some nodes are more connected with each other than with the rest of the nodes. This is probably because even if we consider all documents that are related to “San Francisco,” there are still actually a wide range of documents related to different aspects of “San Francisco.” If we look at the network of tags, we can see that tags related to “San Francisco” include *food*, *travel* and *culture*. Thus, these smaller clusters probably correspond to documents with more specific topics. More analysis will be performed in the future to verify this hypothesis.

## 5.3 Identifying the Topics of Documents

There are some documents (rectangular nodes in the network) which we cannot classify them into either the category of “science fiction” or “San Francisco.” This is because either the documents are only very loosely related to one of these topics, or the tags associated with it are not indicative enough. However, as these rectangular nodes are located in one of the clusters we have observed, it becomes possible to judge, with high probability, the topics of these documents. Also, folksonomies reflect the classification scheme evolving from the collaborative effort of users. Hence, this judgement is not necessarily aligned with the intention of the author of the document. Rather, by saying that a document is related to a certain topic as judged by its location in the network, we are reflecting the opinions of the users. Thus, by constructing and examining the networks of documents, we are able to place the documents into the appropriate context, allowing us to understand what it is about from the viewpoint of users.

## 5.4 Related Works

Research on folksonomies mainly focuses on relations between tags instead of the semantics of individual tags. For example, Begelman et al. [2] propose an

automatic tag clustering algorithm to tackle the problem of synonyms. A more comprehensive method proposed by [14] is able to discover four different kinds of relations – relevant, conflicting, synonymous and unrelated – between tags. Mika [10] proposes to generate lightweight ontologies which are more meaningful by examining tag relations in the social context instead of studying their co-occurrences in documents. One piece of work which is closely related to topic presented here is that by Wu et al. [18], in which the authors investigate how emergent semantics can be derived from folksonomies. 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.

## 6 Conclusions and Future Work

Our study shows that mutual contextualization does occur among the three basic elements in a folksonomy, and that it is possible to acquire a better understanding of the semantics of ambiguous tags by constructing and studying the networks of documents and users associated with the tag.

Currently, many research works focus on how tagging data in folksonomies can be utilized to provide other services, such as identifying user interests, recommending relevant documents or constructing light-weight ontologies. However, all these applications require a better understanding of the semantics of tags in order to provide accurate and useful results. For example, it would not be wise to match users based on the tags they used without knowing that tags may possess different meanings. Hence, the work presented here can be considered as a first step to acquire a better understanding of folksonomies.

However, challenge remains in that while we can identify different groups of users and documents which correspond to different usage of an ambiguous tag, we still need other methods to integrate these different pieces of information to acquire the full picture. For example, how can we know, without examining every documents, which groups of users and documents are associated with a particular sense of a tag? This will be further investigated in our future work.

Specifically, in the future we will apply our method on other ambiguous tags to observe its performance. We hope to gain more insight on how to devise some automatic algorithms to perform tag meaning disambiguation. We will also study different methods of hierarchical clustering or community-discovering algorithms [4, 11], and investigate how these techniques can be applied to discover clusters of documents and users. It is hope that, by further examining the tags associated with different clusters, we can discover the different senses of a tag, probably by examining the tags being used most frequently in the clusters. Finally, we will extend our study to users as well as documents, and investigate how analysis on tripartite graphs can help discover useful information such as communities of users or clusters of documents with similar topics, which will be very useful in applications such as Web page recommendation or social network analysis.

## References

1. Mathes Adam. Folksonomies - cooperative classification and communication through shared metadata. <http://www.adammathes.com/academic/computer-mediated-communication/folksonomies.html>, 2004.
2. Grigory Begelman, Philipp Keller, and Frank Smadja. Automated tag clustering: Improving search and exploration in the tag space. In *Collaborative Web Tagging Workshop at WWW2006, Edinburgh, Scotland*, 2006.
3. 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.
4. Michelle Girvan and M. E. J. Newman. Community structure in social and biological networks. *PROC.NATL.ACAD.SCI.USA*, 99:7821, 2002.
5. Thomas Gruber. Ontology of folksonomy: A mash-up of apples and oranges. <http://tomgruber.org/writing/mts05-ontology-of-folksonomy.htm>, 2005.
6. T. Hammond, T. Hannay, B. Lund, and J. Scott. Social bookmarking tools (i): A general review. *D-Lib Magazine*, 11(4), April 2005.
7. 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 *Lecture Notes in Computer Science*, pages 411–426. Springer, June 2006.
8. T. Kamada and S. Kawai. An algorithm for drawing general undirected graphs. *Inf. Process. Lett.*, 31(1):7–15, 1989.
9. Cameron Marlow, Mor Naaman, Danah Boyd, and Marc Davis. Ht06, tagging paper, taxonomy, flickr, academic article, to read. In *HYPertext '06: Proceedings of the seventeenth conference on Hypertext and hypermedia*, pages 31–40, New York, NY, USA, 2006.
10. Peter Mika. Ontologies are us: A unified model of social networks and semantics. In *International Semantic Web Conference*, pages 522–536, 2005.
11. M.E.J. Newman. Analysis of weighted networks. *Physical Review E*, 70:056131, 2004.
12. Richard Newman. Tag ontology design. <http://www.holygoat.co.uk/projects/tags/>, 2004.
13. S. Niwa, Takuo Doi, and S. Honiden. Web page recommender system based on folksonomy mining for itng'06 submissions. In *ITNG 2006. Third International Conference on Information Technology: New Generations*, pages 388–393, 2006.
14. 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*, January 2007.
15. Emanuele Quintarelli. Folksonomies: power to the people. ISKO Italy-UniMIB meeting, June 2005.
16. G. Smith. Atomiq: Folksonomy: Social classification. [http://atomiq.org/archives/2004/08/folksonomy\\_social\\_classification.html](http://atomiq.org/archives/2004/08/folksonomy_social_classification.html), 2004.
17. Harris Wu, Mohammad Zubair, and Kurt Maly. Harvesting social knowledge from folksonomies. In *HYPertext '06: Proceedings of the seventeenth conference on Hypertext and hypermedia*, pages 111–114, New York, NY, USA, 2006. ACM Press.
18. 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.