Analysis and Representation of Tagging Practices in Online Communities

Haklae Kim, Simon Scerri, John G. Breslin, Stefan Decker
Digital Enterprise Research Institute
National University of Ireland, Galway
IDA Business Park, Newcastle Road
Galway, Ireland
{first name.last name}@deri.org

Honggee Kim
Biomedical Knowledge Engineering Lab
Seoul National University
28-22 Yeonkun-Dong, Chongno-Ku
Seoul 110-749, Korea
hgkim@snu.ac.kr

Abstract

In this paper we analyse a social structure of an online community defined through tagging practices, and investigate whether useful knowledge about the evolution of social networks can be mined through tagging practices, and whether a more representative social structure has any influence on the tagging experience itself. The results from tagging behaviors also need to be represented semantically, along data pertaining to the social structure in order to support data reuse and integration. We then propose a solution for tag data representation which allows data reuse across different tagging systems. We also propose and discuss the enhancement of FOAF via other RDF vocabularies to reflect social tagging practices.

Introduction

The increasing impact of tagging applications for organizing and sharing of resources is motivating new research that is emerging as a result of the marriage between computer science and social science. A variety of research has been done on these topics: investigating tagging behaviors and practices (Farooq et al. 2007; Golder and Huberman 2006; Halpin, Robu, and Shepherd 2007; Hotho et al. 2006; Marlow et al. 2006); discovering community structure and patterns via network analysis (Cattuto et al. 2007a; 2007b; Golbeck and Rothstein 2008; Mika 2007; Mislove et al. 2007) ; harvesting social knowledge from the results of social tagging (Geyer et al. 2008; Heymann and Garcia-Molina 2006; Li, Guo, and Zhao 2008; Penev and Wong 2008; Sigurbjörnsson and van Zwol 2008). These efforts show that social tagging has become useful to reveal social actions from online communities.

Most studies tend to carry out the analysis using large data sets collected from social sites such as Delicious, Flickr etc. These are good to reveal the macro view regarding the social phenomena of tagging such as small-world effects (Cattuto et al. 2007b) and collective intelligence of tagging behaviors (Sigurbjörnsson and van Zwol 2008). New relations driven by tagging behaviors are continuously established directly as well as indirectly, and the degree of an existing correspondence is dynamic and changes over time. As a result, relationships between entities (i.e. users, tags, and resources) also change often.

Changes of individual tags or users over time could be useful information in a particular community that is part of a huge complex network. For example, if we want to find out users who are using some tags during a particular time, or if we want to look for a group that is working/interested in the same area, these micro analysis would be helpful. Current studies tend to be ill-equipped to sense and reflect these social dynamics from tagging behaviors. Another problem is that few have attempted to address adequate representation given results from previous studies. There are some methods to represent social networks in uniform ways such as FOAF (Graves, Constabaris, and Brickley 30 April 2007) and XFN. Using these it could be possible to describe links among users, for instance LiveJournal and MySpace expose their social connections with FOAF. Compared to these, representation of tagging behaviors is limited, and tagging practices are not described in an explicit structure and are not easy to re-use and update, even if we get meaningful results from tagging behaviors.

The limitation in terms of representations can be corrected via Semantic Web technologies, by providing more specific ontological terms to represent people, relationships and their behaviours. Once all this information is exposed to machines, intelligent technology can be employed to constantly elicit new knowledge by observing online social practices.

In this paper we will focus on the evolution of networks and the semantic representation of the findings via a particular case study involving an implicit online community - Planet RDF [1]. The individuals in Planet RDF form part of an implicit online community, where people who know each with varying degrees actively participate towards a common goal for the exchange of data related to a particular interest.

Data Setup and Methods

Data Collection

We collected RSS feeds data from Planet RDF for the period covering January 2004 to the beginning of April 2008. Planet RDF is an aggregation of blogs belonging to 'Semantic Web enthusiasts and hackers' and it is updated on the

[1] http://planetrdf.com/ - Planet RDF aggregates the weblogs of a set of individuals

hour. The selection of blogs was performed by one of the individuals - Dave Beckett, and most of them are very famous and/or active in the Semantic Web area. The list of the blogs are defined in RDF and linked via the `foaf:weblog` property.

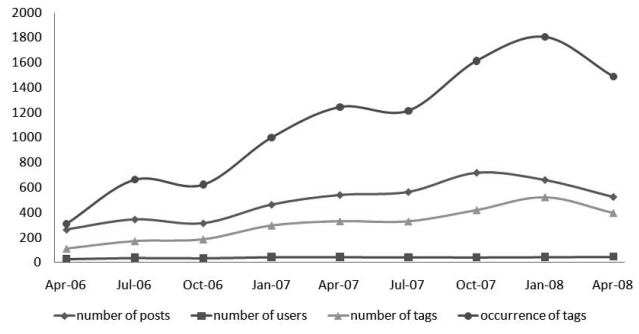

Figure 1: Descriptive Statistics of Data set with Time

There are 58 members in total, including blogs of individual users as well as organizations. The number of members differed according to their level of activity and some members who were included at the beginning of the period were omitted by the end, whereas others had joined. As a consequence, some individuals are not present in all periods. Although most of the published blog feeds were associated with tags, there were some which did not have any.

The posts vary from a low of 261 in April 2006 [2] to a high of 719 in July 2007. The number of users observed in a given period ranges from 28 in April 2006 to 43 in April 2008. The tags assigned by the users are diverse and are not limited only to Semantic Web-related topics. In total there were 4,386 posts which were tagged a total of 9,956 times with 2,750 unique tags. On average, each user published 12.6 posts with 1.7 associated tags every three months. The quarterly average number of tags assigned by a user ranges from a low of 3.9 in 2006 to a high of 12.7 in January 2008. Figure 1 illustrates the number of tags, posts, and members for each of the given periods.

Social Network Analysis

As we stated above, this study aims to gain a first understanding of the relationships between users and tags in Planet RDF and to investigate the possibility of evolving the social structure underlying the users' tagging activities. Network analysis focuses on the relations among entities in a social structure. The network analytic perspective enables us to investigate the characteristics of tagging behaviors within the implicit online community. In our case, the entities consisted of a set of users and a set of tags collected from Planet RDF's RSS feeds. Strictly speaking, tagging is only a personal activity, but if tagging data is shared within a community it can be considered as an emergent collaborative activity, as evidenced in potential social interaction. Moreover, we are also interested in the dynamics of these collaborative activities over time, and its possible influence on the social structure of the participating individuals.

There were a number of possible approaches that we could adopt to analyze this network. In our analysis we focus upon individual properties of a network, in particular *Centrality Measures*. Centrality is a structural attribute of nodes in a network and there are two major centrality measures: *Degree* and *Betweenness* Centrality. In our network analysis we used UCINET to analyse the data and NetDraw [3] to produce the network graphs and also transform the two-mode network into a pair of one-mode networks, one for the users $\mathcal{A}_{user}$ and one for the tags $\mathcal{A}_{tag}$.

Experiments

In this section, we will discuss the findings of our social network analysis applied with respect to the tagging practices in Planet RDF. We will first have a look at the dynamics of tag usage and popularity over time, followed by observations regarding the changing level of the individual users' tagging activities and their level of influence in the community.

Tag Usage Dynamics

The popularity of the 2,750 tags used by the community in the given period varied, although there some trends were clearly visible. '*semantic web*' retained the highest popularity throughout and technology-oriented tags such as '*python*', '*xml*', or '*web services*' also demonstrated a consistent high ranking in early periods. The number of active (i.e. used) tags in the community increased exponentially with time. However we are not just interested in the popularity of tags. We are also interested in which tags are related to most other tags. We can observe these characteristics via centrality measures.

Via a longitudinal analysis, '*rdf*', '*web 2.0*', '*sparql*', '*general*', and '*web*' are the most central tags in this set of tags during the given periods. Whereas top tags in 2006 are technology-oriented, specifically to Internet technologies, after 2006 the domain of the tags becomes broader. More specifically, in 2007 there is a surge in tags related to social technologies, such as '*social software*', '*social media*' and '*social network*'. This reflects the increase in popularity for social software and technologies at the time, also reflected by an increase in posts related to the 'Social Web' domain which became popular in 2007. Similarly 2008 sees the introduction of '*data portability*' whereas the centrality of '*sioc*', '*foaf*', and '*openid*' increases significantly. Thus we can argue that the introduction of the data portability 'topic' indirectly increased the importance of other directly related 'topics'.

During the period of this analysis (2006 to 2008). '*semantic web*' retains the highest ranking on average with regard to both types of centralities during the periods (i.e. the average degree centrality measure C_D^* = 5.44 and the average betweenness centrality measure C_B^* = 30.8). These values are considerably higher than the next most central tags in second and third position with C_D^* = 2.5, C_B^* = 8 for the *web*

[2] We did not consider the data before this period because of lack of data.

[3] http://www.analytictech.com/Netdraw/netdraw.htm

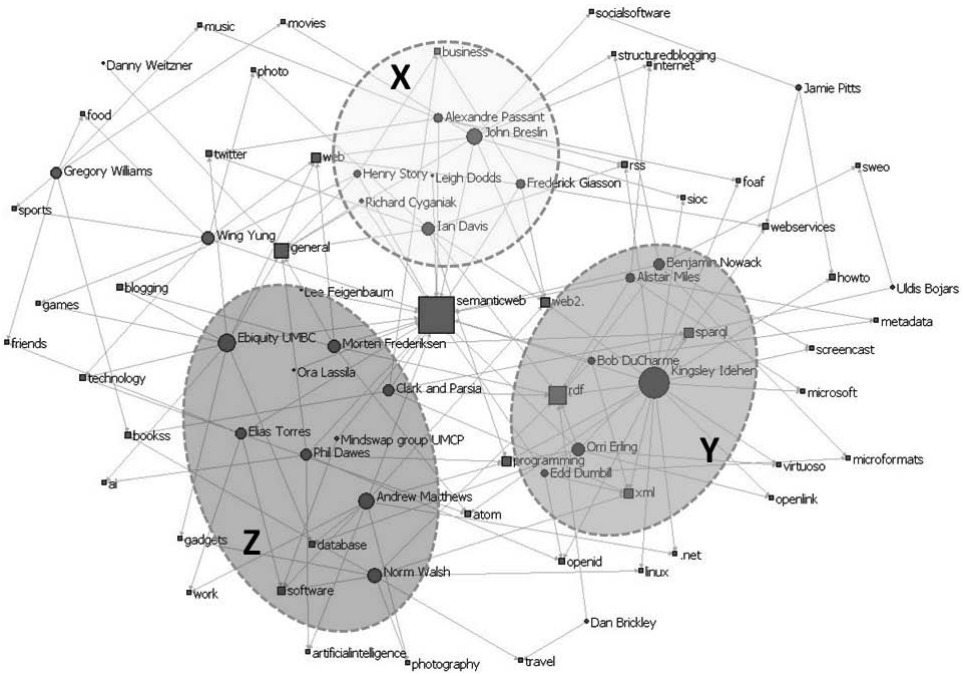

Figure 2: Tag-based Network in April 2007

2.0 and C_D^* = 2.4, C_B^* = 2.9 for the *rdf*. The remaining tags have smaller scores. In particular, some tags in July 2006 have relatively high degree scores, but this is not the case for betweenness centrality. This hints out that in that period tags such as '*web2.0*', '*rdf*', '*data space*' and '*sparql*' are highly used, but singularly, i.e. these tags were not used together, thus their betweenness measures comparatively low than their degree centrality.

User Activity Dynamics

A high ranking of individuals suggests an intermediary role between others. The measures of betweenness roughly match those of degree centrality. However, the rank order has changed significantly over time. For instance, although Idehen, Story, and Breslin are the most active individuals in the network their role in the network changed constantly. Idehen is the most central position in both centrality graphs, because he is the most active (i.e. C_D^* = 8.60, C_B^* = 8.28) in 2006 but also because he uses popular tags. Whereas Breslin and Brickley take the top spots after July 2007.

Barstow and Lee were active in the first quarter of 2006 (C_D (Barstow) = 6.5, C_D (Lee) = 6.2), but their degree centrality plummetted after the beginning of 2007. In contrast, new individuals gain centrality e.g. ebiquity [4], Powers, Passant, and Breslin.

Although some individuals do not have a general high degree index, they have comparatively high betweenness indices in particular periods. For instance, Brickley, Dawes, Davis and Dumbill have a considerably high betweenness ranking. This suggests that although these individuals might have not been very active in the community, the tags they did use played an important role in weaving the network, because they bridge them with individuals from different network clusters.

Tag-based network over time

We will now perform a more visual analysis to learn about the evolution of the community with regards to the tagging practices taking place within. A tag-based social network refers to the social structure that is implied by tagging and the choice of tags (e.g. people are connected when they use the same set of tags).

We will contrast graphs for the tag-based networks during two particular periods, April 2007 and April 2008, to observe whether specific groups of individuals tend to share a number of tags and whether they 'stick together' in a 'common-interest' cluster over extended periods of time. In order to do this we provide the tag-based network as of April 2007, shown in Figure 2. Hence we can directly contrast the tag-based network of the community and its dynamics over a one year period. In 2007, we see that there are three clusters (shaded areas X-Z) based on their tag usage. At first glance one can see that the network in general went through a considerable change. We outlined four user clusters present in this period (shaded areas A-D). Clusters are formed via a number of *Defining Tags*, e.g. the defining tags for cluster A are *owl, web, database, ai, general* and *semantic web*. Clusters are bridged via *Bridging Tags* - tags that are popular with individual users within multiple clusters, e.g. the tags *semantic web, foaf* and *rdf* weaves all four clusters together.

[4] In contrast to other entities, ebiquity represents an organization - ebiquity research group UMBCbiquity research group UMBC

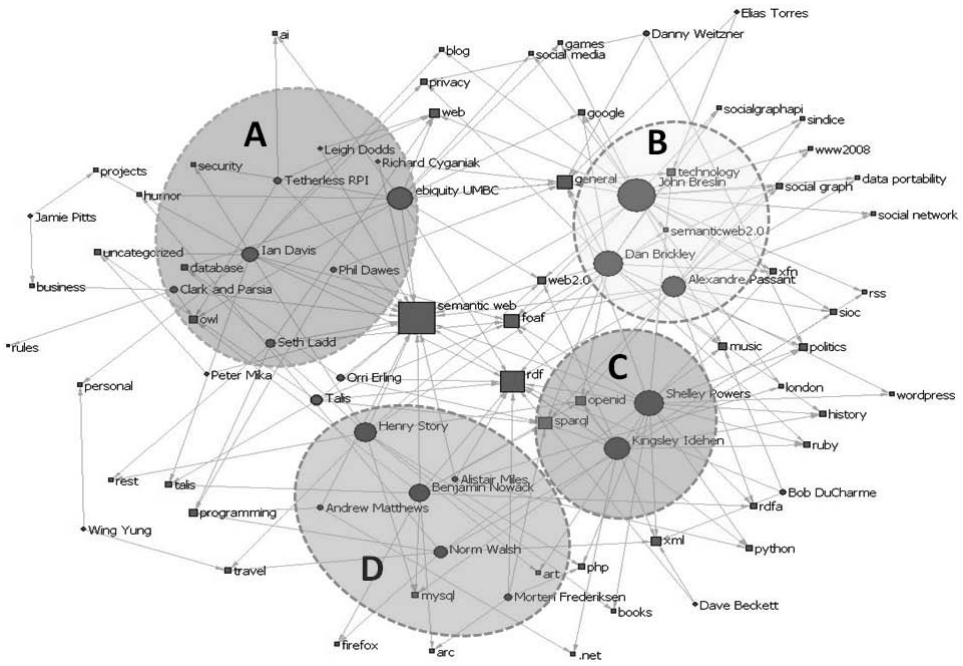

Figure 3: Tag-based Network in April 2008

However, a bridging object does not necessarily need to be popular with a majority of cluster users, e.g. *general* is popular only with a minority of users in clusters A, B and C; yet it brings them together. Inter-cluster tags tend to represent very general concepts, e.g. *programming*, *web*, *web 2.0*, *general*, *twitter*, *internet* and *rss*.

We see that there was a higher use of tags in 2008, and more of these tags had a bridging role between clusters (contrast to the betweenness graph for the periods in question). Considering the individual clusters we see that there are some similarities in between. In particular Passant and Breslin are included within both cluster X (2007) and cluster B (2008). Cluster X was defined by the following tags: *semantic web, web 2.0. web, rss, twitter, business*; whereas B was defined by *sioc*, *social graph, www2008, rdf, sparql, google, foaf* and *openid* amongst others. This suggests that while the focus of cluster X changed, it still maintained its members as seen in B and also included a new user (i.e. Brickley). Another interesting similarity can be observed between cluster Y (2007) and clusters C and D (2008). Nowack and Miles are within both Y and D whereas Idehen and DuCharme are in clusters Y and C. By analyzing the defining tags of C, D and Y we can deduce whether C and D are the result of a split in cluster Y. We found that whereas some of Y's defining tags disappeared by 2008 (e.g. *virtuoso, microformats*), some of them are indeed bridging tags for C and D (e.g. *sparql*, *opened*, *xml*). Thus it is possible that the increased participation in the online community lead to clusters in 2007 breaking down into multiple interest clusters. However we also note that two of the bridging tags for C and D also serve as a bridge between these two clusters and B (*openid* and *sparql*).

Representing tagging behaviors

To enable the results described above, the representation must account for the full environment of tagging. A tagging event is comprised of a resource, a tag, a tagger, and a date, as defined in Newman's ontology that aims to model the relationships between tagging entities. In the case of some properties (i.e. tag:associatedTag, tag:taggedBy, taggedOn, and tag:taggedResource), the relationships focus on a single tagging action and there is no mechanism to describe collective functions for tagging entities. In general, a user or community may have a number of tagging events with arbitrary relationships in between. As the user continues his/her tagging activity, the relationships of tagging entities (e.g. occurrence of tags, co-occurring tags, etc.) should then be updated. This kind of scenario can be achieved by using the SCOT ontology. This ontology extends Newman's and introduces some approaches defining collective and aggregative properties of tagging activities. For instance, a single tagging event is represented in scot:TagCloud. This class provides for metadata related to tagging activities, connecting basic entities such as users, tags, and resources. The scot:taggingActivity describes a relationship between scot:TagCloud and tag:Tagging. Thus, all tagging events for a user are collectively linked to the TagCloud class. Multiple tags in tagging events are aggregated to one unique scot:Tag, if the names of the tags coincide. At the same time, occurrences of the tags are updated via two properties: scot:ownAFrequency and scot:ownRFrequency. The scot:lastUsed and tag:taggedOn properties provide for meta-

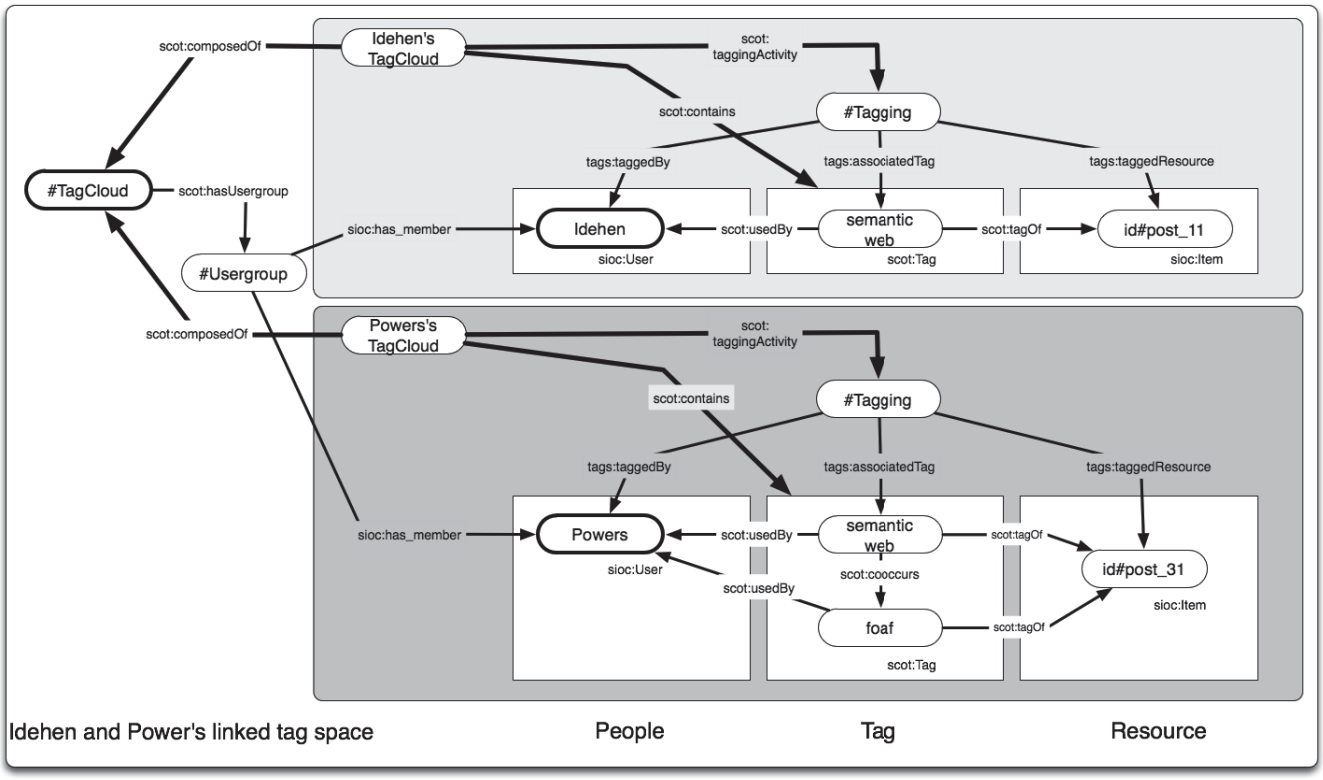

Figure 4: Representing tagging practices (April 2008) of both Idehen and Powers using SCOT

data related to time. scot:Tag is linked to scot:TagCloud via the scot:contains property. Tagger (user) information is represented using SIOC (Bojārs et al. 2008). SCOT also introduces the scot:taggingAccount describing the relationships between a tagging activity and the account used when performing the tagging. The TagCloud class can describe metadata information such as when the tagging occurred (dcterms:created), where the tagging occurred (scot:tagSpace), and how many tags (scot:totalTags) posts (scot:totalPosts) have. Individual tagging events influence all these properties and the TagCloud class plays an important role to define relationships amongst tagging entities. Through this perspective, these representations can be considered as semantic representations of personal folksonomies.

The challenge arises when faced with the problem of describing collective tagging activities for multiple users or communities. As we denoted in Figure 3 , Idehen and Powers shared many tags in April 2008 and their tagging information can be modeled as personal folksonomies via SCOT as shown above. To achieve this, SCOT introduced the scot:composedOf property to link multiple TagClouds. With this property, cluster C in Figure 3 can be described using SCOT instances as illustrated in Figure 4. The tagging information of both users can be stored within one tag cloud and simultaneously interlinked between them.

The SCOT project [5] proposes some applications to generate SCOT metadata and manage, share, and search this data on the Web.

From a social network perspective, SCOT is limited to be able to fully represent user profiles. Thus, we propose an extended solution by interlinking SCOT with existing vocabularies to address their limitations with respect to representing tagging data (Kim et al. 2008). FOAF enables users to describe a set of tags in their profiles via `skos : concept`. However this method does not enable users to link multiple sets of tags generated through heterogenous sources. For instance if a user participates in Delicious and Flickr, it is not easy to link their tagging data. Tagging data can possibly be included as part of a user's FOAF profile. This is possible through SCOT - a SCOT instance (i.e a set of tags) can be linked via `foaf : interest` to indicate an 'interest' of a `foaf : Agent`. This method supports data portability since decentralised tagging data could in this way be reused. We propose the development of recommended models to describe how tagging activities can be connected to people, for example, by using a tagCloud property to connect a TagCloud to the `foaf : Person` or `sioc : User` who made it.

[5] http://scot-project.org

Conclusion

There are many other user-generated practices which one can target to mine hidden social structures, such as posting comments, track-backs, rating, or blogrolls etc. In this paper we considered tagging and we analysed a small and well-organized community - PlanetRDF - to reveal hidden structure from tagging behaviors of users. As we stated in the beginning of the paper, tagging practices in the community are quite dynamic, even if this community is relatively small. Via longitudinal analysis, we found out that interests of the community have changed and this is influenced by the relationships amongst users. It will be useful to have a system where the social networks of users are updated directly or indirectly via suggestions given their tagging practices. For this purpose, interlinking RDF vocabularies (e.g. FOAF, SIOC and SCOT) can be useful to represent the user's interest network as well as their social network.

Acknowledgments

This material was supported in part by Science Foundation Ireland under Grant No. SFI/02/CE1/I131 and by MKE & IITA through IT Leading R&D Support Project.

References

Bojārs, U.; Breslin, J. G.; Peristeras, V.; Tummarello, G.; and Decker, S. 2008. Interlinking the social web with semantics. *IEEE Intelligent Systems* 23(3):29–40.

Cattuto, C.; Baldassarri, A.; Servedio, V. D. P.; and Loreto, V. 2007a. Emergent community structure in social tagging systems. In *Proceedings of the European Confeence on Complex Systems*.

Cattuto, C.; Schmitz, C.; Baldassarri, A.; Servedio, V. D. P.; Loreto, V.; Hotho, A.; Grahl, M.; and Stumme, G. 2007b. Network properties of folksonomies. *AI Communications* 20(4):245 – 262.

Farooq, U.; Kannampallil, T. G.; Song, Y.; Ganoe, C. H.; Carroll, J. M.; and Giles, L. 2007. Evaluating tagging behavior in social bookmarking systems: metrics and design heuristics. In *GROUP '07: Proceedings of the 2007 international ACM conference on Supporting group work*, 351–360. New York, NY, USA: ACM.

Geyer, W.; Dugan, C.; DiMicco, J.; Millen, D. R.; Brownholtz, B.; and Muller, M. 2008. Use and reuse of shared lists as a social content type. In *CHI '08: Proceeding of the twenty-sixth annual SIGCHI conference on Human factors in computing systems*, 1545–1554. New York, NY, USA: ACM.

Golbeck, J., and Rothstein, M. 2008. Linking social networks on the web with foaf: A semantic web case study. In Fox, D., and Gomes, C. P., eds., *AAAI*, 1138–1143. AAAI Press.

Golder, S., and Huberman, B. A. 2006. The structure of collaborative tagging systems. *Journal of Information Sciences* 32(2):198–208.

Graves, M.; Constabaris, A.; and Brickley, D. 30 April 2007. Foaf: Connecting people on the semantic web. *Cataloging Classification Quarterly* 43:191–202(12).

Halpin, H.; Robu, V.; and Shepherd, H. 2007. The complex dynamics of collaborative tagging. In *WWW '07: Proceedings of the 16th international conference on World Wide Web*, 211–220. New York, NY, USA: ACM.

Heymann, P., and Garcia-Molina, H. 2006. Collaborative creation of communal hierarchical taxonomies in social tagging systems. Technical report, Computer Science Department, Standford University.

Hotho, A.; Jschke, R.; Schmitz, C.; and Stumme, G. 2006. Trend detection in folksonomies. In *Proc. First International Conference on Semantics And Digital Media Technology (SAMT)*, volume 4306 of *LNCS*, 56–70. Springer.

Kim, H. L.; Passant, A.; Scerri, S.; Breslin, J. G.; and Decker, S. 2008. Review and alignment of tag ontologies for semantically-linked data in collaborative tagging spaces. In *Proceedings of the Second IEEE International Conference on Semantic Computing 2008*.

Li, X.; Guo, L.; and Zhao, Y. E. 2008. Tag-based social interest discovery. In *WWW '08: Proceeding of the 17th international conference on World Wide Web*, 675–684. New York, NY, USA: ACM.

Marlow, C.; Naaman, M.; Boyd, D.; and Davis, M. 2006. Ht06, tagging paper, taxonomy, flickr, academic article, to read. In *HYPERTEXT '06: Proceedings of the seventeenth conference on Hypertext and hypermedia*, 31–40. New York, NY, USA: ACM.

Mika, P. 2007. Ontologies are us: A unified model of social networks and semantics. *J. Web Sem.* 5(1):5–15.

Mislove, A.; Marcon, M.; Gummadi, K. P.; Druschel, P.; and Bhattacharjee, B. 2007. Measurement and analysis of online social networks. In *Internet Measurement Comference*, 29–42. ACM.

Penev, A., and Wong, R. K. 2008. Finding similar pages in a social tagging repository. In *WWW '08: Proceeding of the 17th international conference on World Wide Web*, 1091–1092. New York, NY, USA: ACM.

Sigurbjörnsson, B., and van Zwol, R. 2008. Flickr tag recommendation based on collective knowledge. In *WWW '08: Proceeding of the 17th international conference on World Wide Web*, 327–336. New York, NY, USA: ACM.