Dynamic Features of Social Tagging Vocabulary: Delicious, Flickr and YouTube

Daifeng Li[1], Ying Ding[2], Zheng Qin[1]
[1]School of Information management and engineering
Shanghai University of Finance and Economics
Shanghai, China
ldf3824@yahoo.com.cn, qinzheng@mail.shufe.edu.cn
dongtianxi@hotmail.com

Staša Milojević[2], Bing He[2], Erjia Yan[2], Tianxi Dong[1]
[2]School of Library and Information Science
Indiana University
Bloomington, IN, USA
{dingying, smilojev, binghe, erjia} @Indiana.edu

Abstract— **This article investigates the dynamic features of social tagging vocabularies in Delicious, Flickr and YouTube from 2003 to 2008. It analyzes the evolution of the usage of the most popular tags in each of these three social networks. We find that for different tagging systems, the dynamic features reflect different cognitive processes. At the macro level, the tag growth obeys power-law distribution for all three tagging systems with exponents lower than one. At the micro level, the tag growth of popular resources in all three tagging systems follows a similar power-law distribution. Moreover, we find that the exponents of tag growth varied in different evolving stages of popular individual resources.**

Keywords-social tagging; dynamic feature; social vocabulary

I. INTRODUCTION

Tagging is a primary means for adding metadata to resources in Web 2.0 environment. Shirky claimed that social tagging reflects the vocabulary and conceptual associations of users [13]. This collective vocabulary speaks the same language as the users and reflects their interests. With their uncontrolled nature and organic growth, user-generated vocabularies have the ability to adapt quickly to changes in both the needs and the vocabulary of users. Nowadays, people's life can be divided into physical life in which people lives and works in concrete feasible places, and virtual life in which people mainly "live" on the Web. These two lives both develop their own language. The language for physical life is the current language that people speak every day, while virtual language (we called it social language here) is what they speak on the Web.

Among the various running social tagging systems, Delicious, Flickr and Youtube are most popular and most studied ones. On Delicious, users usually store and tag bookmarks for future retrieval. On YouTube, instead, videos are mostly tagged by users who uploaded them. Flickr contains user-contributed resources, and tagging rights are divided to permission-based tagging, self-tagging, etc. instead of a free-for-all.

Analyzing the features of the social language can help us understand the statistical characteristics of tagging systems, social relationship of users, and semantic associations among tags. However, it is not well understood what the macro (i.e., the social tagging system taken as a whole) and micro (i.e, individual resources in one social tagging system) features of the social language are, nor how this language evolves. In this study we aim to detect some of the dynamic features of this evolving social language/vocabulary in Delicious, Flickr, and YouTube. The enhanced understanding of macro and micro features of the social vocabulary can be utilized in future application. This paper is organized as follows. Section II introduces the related work. Section III provides the methodology. Section IV discusses the results. Section V evaluates our findings and compares them with others. Section VI presents our conclusions.

II. RELATED WORK

Social tagging activities consist of three major components: tag, tagger and object. For example, tagger A tags object B with tag C. Co-occurrence based clustering remains one of the dominant approaches to identify the relations among tags, taggers and objects [1, 2] so as to use them for recommender systems [5] and folksonomy forming [10], to name a few. However, few of them further explored the dynamics of tagging behaviors and their social vocabulary.

Halpin et al. [7] analyzed dynamics of collaborative tagging system by focusing on the "short head" rather than the "long tail", combined with measures of stability of tag frequencies and information value (the measure of a tag based on the number of pages it retrieves). Schmitz et al. [12] studied the network structure of Bibsonomy and Delicious and found small world network characteristics. They looked at relative path lengths across the tripartite network, and identified hierarchical structures in the network of tag co-occurrence. Cattuto et al. [1] analyzed a large-scale Delicious tagging data to understand the growth of different tags in this system. They studied the temporal evolution of the global vocabulary sizes. As a result, they identified the power-law behaviors and found that the observed growth follows normal distribution throughout the entire history of Delicious and across very different resources. Damianos et al. [3] conducted a statistical analysis of dynamic features of social tagging activities and identified features of social influences and behavioral evolution. Golder and Huberman [6] studied how certain users' sets of distinct tags continue to grow linearly as new resources are added. Other

This research is supported by China National Natural Science Foundation (70971083) and Graduate Innovation Fund of Shanghai University of Finance and Economics (CXJJ-2008-330)

researchers applied tag dynamic features to create recommender systems and make prediction [9, 17].

These studies provide a good starting point for us to understand the characteristics of different social tagging systems. However, few conducted detailed analysis of the evolution of social vocabularies and looking for reasons of causing macro dynamic features from the hidden patterns of individual resources. In this paper, we will address these issues and identify reasons of macro features of social vocabulary from individual resources. We conduct social tagging analysis in the three systems with substantial large coverage of time span (from 2003-2008). Comparing with previous dynamic social tagging studies, our contributions are: 1) our data coverage not only included Delicious (with longer time coverage) but also extended to Flickr and YouTube to ensure an extensive coverage of social tagging data; 2) we extended the approach to analyze how tags growth of popular resources evolve according to physical time; 3) we studied the dynamic features of tag vocabulary growth for individual popular resources.

III. Methodology

A. Data Collection

We developed a tag crawler based on the Upper Tag Ontology (UTO) to harvest, integrate and store in RDF triples tagging data from Delicious, Flickr and YouTube [4]. To avoid timeouts and to make efficient use of available internet bandwidth, the UTO crawler uses the Smart and Simple Web crawler framework, a multi-thread crawler designed by Torunski [15]. In November 2008, The UTO crawler was used to retrieve tagging data from Delicious, Flickr and YouTube. The crawler identified objects, taggers, tags, dates, comments and votes. In total, the data retrieved contains approximately 3M bookmarks, 0.6M taggers and 15.7M tags harvested from Delicious; 1.4M photos, 0.07M taggers and 17.7M tags harvested from Flickr; and 1.4M videos, 0.8M taggers and 11.3M tags harvested from YouTube.

B. Data processing

The crawled dataset covers Delicious from 2003 to 2008, Flickr from 2004 to 2008, and YouTube from 2005 to 2008. We used unified format {tagger, link, tag{tag1, tag2, tag3…tagk}, time} to represent one post. A post is a tagging event in which one tagger tags one object with several tags. To further process the data, we deleted those data which existed before the system was established (e.g., there are some tags in Delicious appeared before 2003), posts with missing values (such as, no tagger, no link, no tag, or no date), and repeat annotation activities of taggers (for example, a tagger may bookmark the same link with the same tag more than once).

C. Experimental Data

After data processing, we obtained 3,006,706 posts from Delicious, 1,380,734 posts from Flickr, and 1,372,315 posts from YouTube. Table 1 summarizes the basic statistics regarding the three different tagging systems.

TABLE I. Social Tagging Data

Social Network	Objects	Taggers	Tags	Tag/Object	Tag/Tagger	Object/Tagger
Delicious	3,006,706	596,816	15,707,782	5.22	26.31	5.037
Flickr	1,380,734	75,679	17,797,832	12.89	235.2	18.24
YouTube	1,372,315	793,830	11,331,362	8.26	14.27	1.73
Sum	5,759,755	1,466,325	44,836,976	26.37	275.78	25.01

D. Algorithms

Refer to Cattuto's introduction of tag vocabulary growth [1], we build a dynamic tag growth model based on a time counting variable *tg*. For a social tagging system, we sort all posts by their date in an ascending order and initiate the number of tags as 0. Each time when one post is added, we count the number of tags in that post as m, and update *tg* as *tg=tg+m*. Based on this, we then proposed two main algorithms to evaluate the dynamic features of social tags.

Macro Tag Growth Algorithm (MaTGA): this algorithm calculates the evolution of tags at the macro level which measures the macro features of tags by taking the social tagging system as a whole. It measures the social vocabulary growth *f(tg)* in a certain tagging system as the function of *tg*. *ps* and *qs* are the two query result sets by performing SQL queries mentioned in the algorithm. The description of MaTGA is below:

```
Function [FTG,TG]=MaTGA()
// FTG is an array of f(tg); TG is an array of tg.
int tg=0;
int f(tg)=0;
int m=0;
int i=0;
ps.perform("select all posts in a social tagging database ordered by real time in an ascending order");
while ps.next() {
    i=i+1
    m=count(ps.tag); //m means the number of tags in the current post;
    tg=tg+m;
    TG(i)=tg;
    qs.perform("select post.tag in a social tagging database where post.tag=ps.tag and post.time<=ps.time");
    f(tg)=f(tg)+m-count(qs.tag);
    //m-count (qs.tag) means the number of tags that never appear before;
    FTG(i)=f(tg);
}
end Function
```

Micro Tag Growth Algorithm (MiTGA): MiTGA measures the micro level of a social tagging system. It analyzes the dynamic features of individual resources inside a social tagging system. We called these individual resources target resources. MiTGA is very similar to MaTGA with a slight modification to select all the posts whose resource is target resource. For example, in MiTGA, if the target resource is www.facebook.com, then only posts that bookmarked this resource are collected and analyzed.

The purpose of designing the two algorithms above is to find some intrinsic features of social tagging. Because the evolution of tag vocabulary growth is a cumulative process of new tags; using *tg* takes the basic unit of tagging behavior, a post event, as an accurate, natural and dynamic reflection of the period of people's tagging behaviors, which can efficiently capture full richness of the dynamics of social tagging system. That's why we decide to use *tg* instead of physical time *t*.

IV. Results and Discussion

We compared the dynamic tag features of three systems from both macro and micro perspectives. The results of macro level analysis are discussed in Subsection A and the results of micro analysis in Subsections B and C.

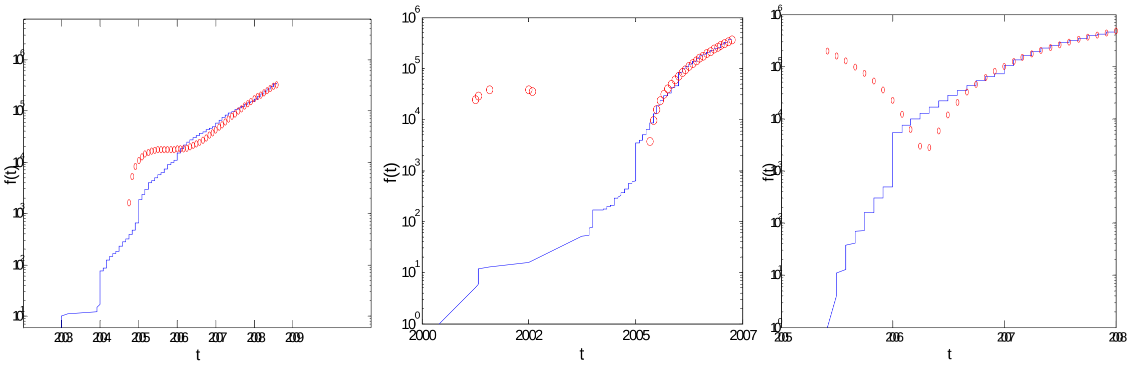

Figure 1. Curve of *f(t)* (in log scale) as the function of physical time *t* in Delicious, Flickr and YouTube

A. Comparsion of macro tag growth in three tagging systems

MaTGA is used to capture the macro dynamic growth of all tags as the function of *tg* in three tagging systems. We can see that all systems closely follow a power-law distribution across the *tg*. The tag growth *f(tg)* satisfies $f(tg) \sim tg^{\gamma}$, where γ is an exponent of power-law distribution (γ represents the increasing rate of tag vocabulary growth). The exponents for Delicious, Flickr and YouTube are $\gamma_{delicious}$ =0.8040, $\gamma_{flicker}$ = 0.8039 and $\gamma_{youtube}$ `= 0.8580 respectively. We can see that for each tagging network, the values of γ are different: $\gamma_{flicker} < \gamma_{delicious} < \gamma_{youtube}$. This can be explained that the social vocabulary of YouTube is more stable than those of Flickr and Delicious; Delicious and Flickr involves more individuals in the collective process through the social functions they provides, resulting in more variations from individual diversities, while videos on YouTube tends to be tagged mostly by the users who uploaded the videos, leading to a more semantically coherent vocabulary.

We also find that γ are very similar in different social tagging systems: in our study, it ranges from 0.8 to 0.9 (in Cattuto's work in Delicious, it is close to 0.8 [1]). While in other systems, such as English corpora, γ ranges from 0.4 to 0.6 [8]; in Thai subset of WWW webpages, it is close to 0.5 [11]. This indicates that the vocabularies of social tagging systems are more stable than normal corpora or the subset of webpages. This further confirms that each social tagging system has a core vocabulary. Other reasons for the difference of γ between social tagging systems and English corpora are: (a) tags are generally nouns and have no grammatical structure, and (b) the number of taggers in social tagging systems is increasing, while the number of authors in English corpora is limited [16].

Figure 1 shows the growth of new tags (presented in logarithmic scale) along with physical time *t* in the three systems. We can find that although the slope changes of each curve present noises at the beginning, they later (we use red spot curve to approximate them in the figures), display the characteristics of exponential function along with physical time *t* (by month). When the stable vocabulary is formed, the tag vocabulary growth comes into a stable evolving status, which displays an exponential function along with physical time *t*. Furthermore, this stable vocabulary contains the frequently used tags. The formation of the vocabulary may take several years (in that period, the curve displays an irregular growth) and it differs in different tagging systems. For example, in Delicious and Flickr, it takes around 5 years to form such stable vocabulary, while it takes 3 years for YouTube.

After the formation of each social vocabulary for three social tagging networks, we find that each curve satisfies an exponential increase with exponent γ<1 (red curves in Fig. 1), suggesting that in most situations taggers tend to use tags that already exist in the social vocabulary to describe resources.

B. Comparision of micro tag growth in three tagging systems

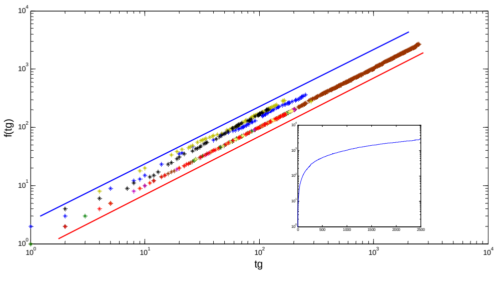

Figure 2. Tag growth for 10 popular resources in Delicious

Cattuto [1] has proved that macro tag growth exponent is similar to micro tag growth exponent of popular resources in Delicious. Micro tag growth means the tag growth of a certain resource. It can be captured by using MiTGA. To compare our findings with Cattuto's [1] on Delicious (from 2003 to 2006), we used a much larger Delicious dataset covering 2003-2008. We selected 10 out of the 1,000 top ranked resources. To make the selection, we took the top-ranked resource and one after every 100th. We draw lines of tag growth as the function of *tg* for each resource (Figure 2).

In Figure 2, the tag growth of all 10 popular resources shows a sub-linear feature with parallel consistent growth after a period of time. They satisfy power-law distribution and their exponents are between 0.5786 and 0.9245 (see the blue line and the red line). Also, the slope of the micro tag growth of each resource decreases along with physical time *t* (shown in the embedded figure) which means the rate of creating new tags becomes smaller and will reach a fixed value after a period of time. This analysis cannot be conducted in Flickr and YouTube because the number of taggers who tagged popular resources is too small. This fact also reflects the differences of intrinsic nature of tagging behavior across the three social tagging systems. In Flicker, there are restrictions of "self-tagging" and "permissive-tagging"; and in YouTube, uploaders account for a dominant part of taggers, so even popular resources are tagged by a relatively small number of users.

As the number of taggers per resource is less than 2 in YouTube, it is hard to calculate the probability distribution of γ_{micro} as majority of resources have similar γ_{micro} values. This again reflects that videos on YouTube are tagged mostly by users who uploaded them, leading to the relatively small number of taggers per resource.

C. *Comparision of tag growth exponent probability distribution for popular, less-popular and non-popular resources in three tagging systems*

The tag growth exponent of a certain resource changes over time, so we can compute its exponent γ_{micro} at the final spot of *tg* by using the formula $\gamma_{micro} = \log(f(tg_{\max}))/\log(tg_{\max})$. According to [1], for a group of popular resources (i.e., top 1,000 ranked resources), the probability distribution of exponent γ_{micro} satisfies a Gaussian distribution, and the mean value is close to the exponent of macro tag growth. For the group of less-popular resources (i.e., top 100,000-101,000 ranked resources), the distribution is not normal. For the group of non-popular resources (i.e., the last 1,000 ranked resources), the possibility of γ_{micro} is random. There are two kinds of non-popular resources: those existing in the systems for quite a long time and never becoming popular, and those that are either new or newly tagged. The probability of γ_{micro} is random, which indicates that very few taggers tag these non-popular resources.

In our Delicious, Flickr, and YouTube datasets, we select three groups of resources according to their ranks in each tagging system: the top 1,000 ranked resources as popular resources, 100,000-101,000 ranked resources as less-popular resources, and the last 1,000 ranked resources as non-popular resources.

In Figure 3, the exponent of tag growth of top 1,000 ranked resources at the final spot of *tg* in Delicious follows normal distribution with its mean value as 0.72, while less-popular and non-popular do not follow normal distribution. In Figure 4, the top 1,000 ranked resources in Flickr have not yet reached the same distribution as top 1,000 ranked resources in Delicious, probably due to the fact that Flickr imposes many restrictions on tagging (e.g., "self-tagging" and "permissive-tagging"). In this case, number of different taggers who tagged those popular resources differs according to exogenous restrictions, resulting in the non-normal distributions of the exponent probability distribution of Flickr popular resources.

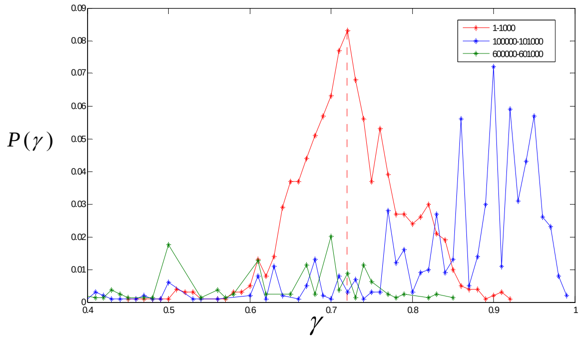

Figure 3. Exponent probability distribution of groups of resources in Delicious

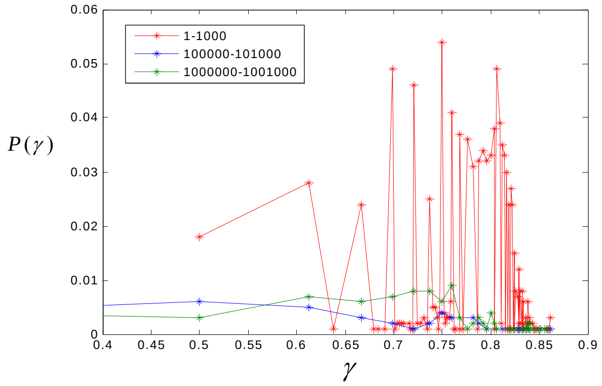

Figure 4. Exponent probability distribution of groups of resources in Flickr

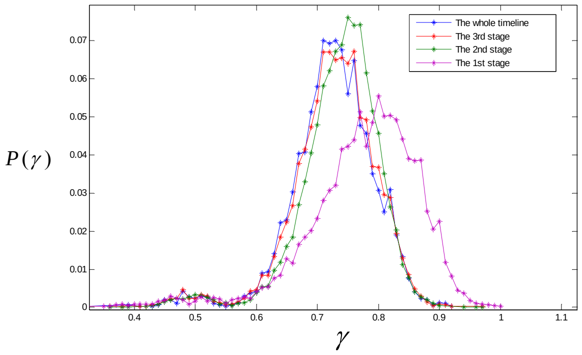

Figure 5. Tag growth exponent probability distribution of resources group in different development stages.

In order to understand the normal distribution of exponent of tag growth in top ranked resources in Delicious, we selected top 5,000 ranked resources and used the same method to compute the exponent probability distribution of each resource. We calculated the timeline for each resource by subtracting its latest date of tagging from its earliest date of tagging and divided this timeline into 4 stages. For each stage, we computed the tag growth exponent probability distribution of the selected 5,000 resources (Figure 5).

Figure 5 provides insight into how popular resources are formed. At the first stage of each popular resource, the exponent probability distribution follows normal distribution (the absolute value of their skewness is 1.0043). Their exponent mean value is 0.81, and the absolute value of kurtosis is 0.5337, which is bigger than that of the 2nd, 3rd and 4th stage. In the next three stages, the absolute values of kurtosis are becoming smaller while physical time and the absolute values of skewness are becoming bigger. The value of exponent tends to be 0.7.

We further tested this on all the Delicious resources and found that the value of γ_{micro} of each resource is around 0.72 (with standard error 0.02). We also found that the proportion of the resources whose γ_{micro} is between 0.7 and 0.74 is around 3.7%. For all the resources in Delicious, if tag growth exponent reaches 0.72, its tag vocabulary approaches a stable status.

V. Evaluation

The macro tag growth of social tagging systems is similar to English corpora and academic articles whose vocabulary growth obeys power law distribution and the exponent has a sub-linearity along with *tg* [1]. Researchers found that the range of macro vocabulary growth exponent of traditional English corpora and academic articles is between 0.4 and 0.6 [8]. We found the exponent range of social tagging systems between 0.8 and 0.9. Social language has no grammatical structure but it contains significant semantic information. The micro tag growth of certain resources is similar to the growth of vocabulary in papers and articles, with both having sub-linearity features along with time. Based on this we can use similar methods to deal with resources in social tagging systems.

Different social tagging systems also have different dynamic features. We used Delicious data (with the addition for 2007-2008) to compare our findings with those of Cattuto et al. [1]. We found that the results are consistent with respect to the growth exponents of macro tags and micro tags. The values of tag growth in Flickr and YouTube were not consistent with the values obtained for Delicious. Our finding also confirms Suchanek, Vojnovic and Gunawardena's finding based on a social tagging analysis of 65,000 Delicious bookmarks and a user study of over 4,000 participants [14]. We both found that the popular resources have more stable tags.

VI. Conclusion

In this paper, we built up a dynamic model to analyze the features of three social tagging systems: Delicious, Flickr and YouTube based on large scale tagging data crawled by the UTO crawler. For the social vocabulary, the macro tag growths in three social tagging systems follow a power-law distribution. There exists a relatively stable vocabulary (mainly consisting of frequently used tags) to describe the content of a social tagging system. It takes 5 years for Delicious and Flickr and 3 years for YouTube to form such stable vocabulary. The social vocabulary of YouTube is more stable than those of Flickr and Delicious. The social vocabulary of tagging systems is more stable comparing with normal corpora or subset of webpages. Comparing with Delicious, the popular resources in Flickr have not reached the popular level of Delicious, while they are roughly at a less-popular status (with fewer unique taggers per resource) of Delicious. For all the resources in Delicious, if the resource's tag growth exponent reaches 0.72, its tag vocabulary becomes stable.

References

[1] C. Cattuto, A. Baldassarri, V. Servedio, and V. Loreto. Vocabulary growth in collaborative tagging systems. http://arxiv.org/abs/0704.3316. Accessed 15th March 2010.

[2] C. Cattuto, D. Benz, A. Hotho, and G. Stumme. Semantic grounding of tag relatedness in social bookmarking systems. In Proceedings of 7th International Semantic Web Conference ISWC2008, vol. 5318, Lecture Notes in Computer Science, Karlsruhe, German, pp. 615–631, 2008.

[3] L. Damianos, J. Griffith and D. Cuomo. Onomi: Social Bookmarking on a Corporate Intranet. In Collaborative Web Tagging Workshop, Proceedings of the 15th International WWW Conference, Edinburgh, Scotland, 2006.

[4] Y. Ding, E. Jacob, M Fried, I. Toma, E. Yan, S. Foo & S. Milojevic. (2010 forthcoming). Upper Tag Ontology (UTO) for integrating social tagging data. Journal of the American Society for Information Science and Technology.

[5] G. I. Fountopoulos. Richtags: A social semantic tagging system. Master thesis, School of Electronics and Computer Science, University of Southampton, 2007.

[6] S. Golder and B. Huberman. The structure of collaborative tagging systems. Journal of Information Sciences, 32:198–208, 2006.

[7] H. Halpin, V. Robu, and H. Shepherd. The complex dynamics of collaborative tagging. In Proceedings of the 16th international WWW Conference, New York, NY, USA, pp. 211–220, 2007.

[8] D. Harman Overview of the Third Text Retrieval Conference. In Proceedings of the 3rd Text REtrieval Conference (TREC-3), NIST Special Publication 500-207, 1–19, 1995.

[9] P. Heymann, D. Ramage, H. Garcia-Molina. Social Tag Prediction. In Proceedings of the 31st annual international ACM SIGIR Conference on Research and development in IR, Singapore, 531-538. 2008.

[10] A. Hotho, R. Jaschke, C. Schmitz, and G. Stumme. Information retrieval in folksonomies: Search and ranking. In Y. Sure and J. Domingue (ed.), The Semantic Web: Research and Applications. Springer, Heidelberg, 2006.

[11] S. Sanguanpong, S. Warangrit, and K. Koht-arsa. Facts about the thai web. http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.11.7453 Accessed 15 March, 2010.

[12] C. Schmitz, M. Grahl, A. Hotho, G. Stumme, C. Cattuto, and A.Baldassarri. Network properties of folksonomies. In Tagging and Metadata for Social Information Organization; a workshop Proceedings of the 16th International WWW Conference, Banff, Alberta, Canada, May 8-12, 2007.

[13] C. Shirky. Ontology is overrated: Categories, links, and tags. http://shirky.com/writings/ontology_overrated.html. Accessed 6 March 2010.

[14] F. M. Suchanek, M. Vojnovic, and D. Gunawardena. Social tags: Meaning and suggestions. In Proceedings of the 17th ACM Conference on Information and Knowledge Management, Napa Valley, California, USA, October 26-30, 2008.

[15] L. Torunski. Smart and simple webcrawler. https://crawler.dev.java.net. Accessed 10 March 2010.

[16] C. Veres. The language of folksonomies: What tags reveal about user classification. In C. Kop, G. Fliedl, H. C. Mayr, and E. M´etais, (ed.) NLDB, vol. 3999, Lecture Notes in Computer Science, Springer, pp. 58–69. 2006.

[17] Z. Xu, Y. Fu, J. Mao, and D. Su. Towards the semantic web: collaborative tag suggestions. In Collaborative Web Tagging Workshop, Proceedings of the 15th International WWW Conference, Edinburgh, 2006.