The Dynamic Features of Delicious, Flickr and YouTub

Nan Lin[1], Daifeng Li[2], Ying Ding[3], Bing He[3], Zheng Qin[2]
, Jie Tang[4], Juanzi Li[4],Tianxi Dong[5]
[1]School of International Business Administration,
Shanghai University of Finance and Economics
Shanghai, China
linn@mail.shufe.edu.cn
[2]School of Information Management and Engineering,
Shanghai University of Finance and Economics
Shanghai, China
ldf3824@yahoo.com.cn
qinzheng@mail.shufe.edu.cn
[3]School of Library and Information Science
Indiana University, Bloomington, IN, USA
{dingying, binghe} @Indiana.edu
[4]Department of Computer Science and Technology,
Tsinghua University, Beijing, China,
jietang@tsinghua.edu.cn
[5]Rawls College of Business,
Texas Tech University, TX, USA.
dongtianxi@hotmail.com

Abstract – **This article investigates the dynamic features of social tagging vocabularies in Delicious, Flickr and YouTube from 2003 to 2008. Three algorithms are designed to study the macro and micro tag growth as well as dynamics of taggers' activities respectively. Moreover, we propose a Tagger Tag Resource LDA (TTR-LDA) model to explore the evolution of topics emerging from those social vocabularies. Our results show that (1) at the macro level, tag growth in all the three tagging systems obeys power-law distribution 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; (2) the exponents of tag growth vary in different evolving stages of resources; (3) the growth of number of taggers associated with different popular resources presents a feature of convergence over time; (4) the active level of taggers has a positive correlation with the macro-tag growth of different tagging systems; and (5) some topics evolve into several sub-topics over time, while others experience relatively stable stages in which their contents do not change much, and certain groups of taggers continue their interests in them.**

Keywords – social tagging, dynamic feature, social vocabulary

1 INTRODUCTION

Social tagging systems such as Delicious, Flickr and YouTube, which have been rapidly gaining popularity on the Internet, allow users to interactively annotate a pool of shared resources using a set of unstructured descriptive terms, called tags, to navigate, browse and retrieve resources. Tagging is a primary means for adding metadata to resources in the Web 2.0 environment and helps to spread ideas, memes, trends and fashions. The act of tagging reflects an individual's conceptual associations and enables loose coordination (Shirky, 2005), but it does not enforce a single interpretation of a tag or a concept.

Social tagging systems have contributed to the formulation of people's online social language. Nowadays, people's lives can be divided into a physical life, where people live and work in concrete places and conduct corporeal activities, and a virtual life, where people "live" on the Web, chatting with their virtual friends within their favorite social networks. Both lives develop their own languages – those learned from parents and in school that are spoken every day, and virtual or online social languages "spoken" on the Web. In order to communicate successfully, to attract attention from social network friends, for example, people need to speak the social languages developed by Internet users, especially social network users.

Underlying the dynamics of taggers' social language is the tagging activity. With their uncontrolled nature and organic growth, user-generated vocabularies have the ability to adapt quickly to changes in both the needs and vocabulary of users. The freedom of natural language, used in these environments of free-will tagging, requires less cognitive effort than making a decision about how well a pre-defined category captures the content of a resource and/or represents the immediate needs of users. The use of existing tags contributed by other users can reflect a consensus emerging from collective tagging, while the proposing of new tags adds to the pool of tags that then go through social filtering, being either co-opted by other users, or largely ignored, thus becoming a singular instance. It is not clear, however, how the proportion of existing and new tags added to one resource over time is related to the age and popularity of resources in social tagging systems.

Furthermore, examining semantic features of tags can crucially add to our insights on this social language. Topics implied by tags can reveal the interests of individual taggers and online communities, as well as the subject content of resources being tagged. More importantly, the

dynamic evolution of the semantic features of tags in these social tagging systems exposes the history of vanishing and merging topics that can be used to predict future trends.

But it is not well understood how this language evolves at the macro level (social tagging systems taken as a whole) and the micro level (individual resources in one social tagging system), nor do we understand how the evolution of social language is determined by tagging activity of individual tags over time. Moreover, the dynamic features of topics revealed by social vocabularies over time also need exploration. In order to advance our knowledge in this area, it is essential to conduct a dynamic study of social vocabularies in social tagging systems from various perspectives and conditions.

Analysis of social tagging systems can open new perspectives for indexing theories, enrich the knowledge of information retrieval, provide ways of approaching the cognitive process of sense-making, and give insight on various interactions among social-technical systems. Tags, also referred to as social tagging vocabularies, are one of the most important carriers of this information. The study presented in this paper is centered on tagging vocabulary. Four different approaches are applied to the study of social tagging vocabulary: 1) modeling growth of social tagging vocabulary at both the macro and micro levels; 2) comparative analysis of active tagger behavior associated with the growth of social tagging vocabulary; 3) modeling the semantic structure underlying social tagging vocabulary using the TTR-LDA model; and 4) by the integration of the different perspectives, common features and unique characteristics of three different major social tagging systems. Therefore, the formation and evolution of social tagging vocabulary are analyzed and unveiled through a multi-dimensional perspective, involving both taggers and tagged objects (i.e., web pages, photos, and videos).

Among the social tagging systems, Delicious, Flickr and YouTube are the most popular. Different design paradigms shape the incentive structures that drive people to tag resources, leading to a diverse range of relationships being expressed by tagging across these three social tagging systems. Delicious is largely task-focused, with a priority on storing bookmarks for future retrieval, and thus organizational motivations are most dominant there. On YouTube only videos are tagged, largely by those users who upload them. Tagging on YouTube primarily exposes the content for discovery by other users, and convergence around conventional meanings can be expected. Flickr contains user-contributed resources, such as

photos taken by users themselves. In Flickr, tagging rights are restricted to self-tagging and permission-based tagging instead of a free-for-all approach (Paolillo, 2008; Marlow, Naaman, Boyd & Davis, 2006).

In this study we identify key dynamic features of this evolving social vocabulary in Delicious, Flickr, and YouTube. Our study enhances the understanding of the growth of social vocabularies at both the macro and micro levels, the dynamics of tagging activities that determine the vocabularies and the evolution of topics revealed by tags.

The contributions of this study include: 1) approaching social tagging vocabulary through multiple perspectives; 2) extending the time counting variable (i.e., post intervals) proposed by Cattuto et al. (2008) from tags to taggers and resources, so as to dynamically reflect the natural settings of social tagging context; 3) Performing various views of vocabulary growth including macro and micro level vocabulary, number of taggers associated with subsets of resources, as well as an increase in the number of tags created by specific taggers; and 4) Conducting a comparative analysis for three of the typical social tagging systems over substantial data coverage.

This paper is organized as follows: Section 2 introduces relevant related work in this area; Section 3 explains the methodology of our study; Section 4 discusses the results; Section 5 evaluates our findings and compares them with others; and Section 6 presents conclusions.

2 RELATED WORK

Social tagging activities take the form of a triple {tag, resource, tagger}, wherein each element is connected to the other through tagging behavior. Since their emergence in the last decade, social tagging systems have received much interest among researchers in various fields. The main trend of studies on social tagging systems has been the identification of relations among tags, taggers and objects through co-occurrence-based clustering (Cattuto, Baldassarri, Servedio, & Loreto, 2007; Cattuto, Benz, Hotho, & Stumme, 2008), which provides implications for recommender systems (Fountopoulos, 2007), personalized search engines (Xu et al., 2008), and folksonomy forming (Hotho et al., 2006), among others. Recent work on social vocabularies generated by taggers has explored their linguistic characteristics (Kipp & Campbell, 2006b), their growth (Golder & Humberman, 2006; Cattuto et al., 2007), their topic structure (Li et al., 2008) and their effectiveness for browsing/searching resources (Morrison,

2008), visualizing trends (Dubinko et al., 2006), identifying patterns in tagging behaviors (Schmitz et al., 2006) and ranking terms in a vocabulary (Harman, 1995).

Many articles in a broad range of disciplines have been devoted to the mechanism that drives the forming and evolving of folksonomy. As proposed by Trant (2009), studies on social tagging and folksonomy focused on three aspects: 1) folksonomy itself and the role of tags in indexing and information retrieval; 2) tagging and the behavior of users; and 3) social tagging systems as technical frameworks. Since the emergence of social tagging systems, people have noticed that it challenges traditional classification schemes and controlled vocabularies-based indexing. Mathes (2004) is the one of the first researchers who reviewed social tagging systems as uncontrolled metadata. No rules of indexing, no professional librarians, and no predefined and well-organized hierarchical subject terms exist in social tagging systems. However, there does exist a collective environment with instant public feedback, registered anonymous Internet users, and free use of natural language. Voss (2007) suggested that tagging should better be seen as a popular form of manual indexing on the Web. He stated that the feedback mechanisms blurred the difference between controlled and free indexing.

In additional to qualitative analysis of social tagging as indexing, some researchers have conducted quantitative comparative analysis on social tagging and traditional indexing. Macgregor and McCulloch (2006) compare folksonomies with controlled vocabularies and found that while social tagging systems have deficiencies originating from the absence of controlled vocabularies, the interactive and social aspects exemplified by collaborative tagging systems, as well as their collective process of information management are beneficial for traditional indexing. Similarly, Kipp (2006a, 2006b, 2007) contrasted tags, author keywords, and professionally supplied descriptors for 176 entries from citeulike.org on a 7-point scale. Results showed that, though related, taggers' terminology differs from that of authors and indexers. Smith (2007) investigated the differences between tags and subject headings in the Library of Congress Catalog. The study suggested that LibraryThing tags outperformed subject headings at identifying identified latent subjects.

Contextualized in a social environment, tagging behavior of users is an intrinsic fundament of social tagging systems. Users' cultural, demographic, and language background will pose systematic inaccuracy in tags they created. Meanwhile, the temporal nature of users' interest

will also be reflected in tags they created. Those research issues have attracted much attention. Marlow, Naaman, boyd, & Davis (2006) provided a model and a taxonomy of tagging systems to frame their analysis and design. Features that were modeled include: tagging rights (what can be tagged by whom); tagging support (tag recommendation); aggregation (removal of duplicate tags); type of object; source of material (originally-created or internet resources); resource connectivity; and social connectivity. Additionally, they categorized users' incentives into organizational and social motivations. Some other studies approach the tagging motivations from the perspective of cognitive science. Tagging can be considered as an act of sensemaking, with shared tags becoming a form of collective consensus. Managing and organizing resources are the most common direct benefit for taggers (Weick et al., 2005).

More focused, tag growth and tagging activity have been studied since social tagging systems first attracted researchers' attention. Golder and Huberman (2006) found that certain user sets of distinct tags continue to grow linearly as new resources are added. Cattuto et al. (2007) analyzed large-scale Delicious tagging dataset to understand the growth of different tags in this system. As a result of studying the temporal evolution of global vocabulary sizes, they identified power-law behaviors of these phenomena and found that the observed growth follows normal distribution throughout the entire history of Delicious and across very different resources. Halpin et al. (2007) analyzed the dynamics of collaborative tagging systems by focusing on the "short head" rather than the long tail, in combination with measures on the stability of tag frequencies and information values (the measure of a tag based on the number of pages it retrieves). They also extended a tripartite model for tagging, using a preferential attachment model which consists of taggers, tags, and resources respectively. Serrano, et al (2009) analyzed the regularity of word growth in Wikipedia, the Industry Sector database (IS) and the Open Directory (ODP). They found that while the number of total words increases, the number of new words also increases, thus satisfying Heap's law (sub-linear features). Serrano, et al also found that the probability distribution of the similarities between pair-wise documents also satisfies Zipf's Law. Altmann, et al (2009) conducted research on the temporal distribution of words in different time intervals. They found that stretched exponential β can be seen as an intrinsic feature of certain words in USENET. The distribution of β can also be seen as an intrinsic feature of semantic classes. Cattuto, et al (2009) found that dynamic co-occurring features of tags fit the power-law distribution with an exponent of around 0.7. The frequency-

rank plot of tags also satisfies the power-law with an exponent of around –1.42. They analyzed the dynamic relationship of degree k, strengths s and weights w of all nodes in the collective network and used the Watts-Strogatz algorithm to generate a random network to compare with the data from Delicious in order to observe the small-world properties of social annotation systems.

Other researchers have applied tag dynamic features to create recommender systems and make predictions (Heymann, Ramage, & Garcia-Molina, 2008; Veres, 2006). Damianos et al. (2006) conducted a statistical analysis of dynamic features of social tagging activities and identified aspects of social influences and behavioral evolution.

As for topic mining of social tags, some pioneer studies applied Latent Dirichlet Allocation (LDA) to social tagging systems. Generally speaking, LDA helps to explain the similarity of data by grouping features of this data into unobserved sets. First introduced by Blei et al. (2003), it was used to solve various tasks such as topic mining (Tang et al., 2008) and community detection (Zhang, et al., 2007). Krestel et al. (2009) applied LDA to recommending tags for resources. Xiance and Maosong (2008) proposed a tag-LDA model, which extends the LDA, model by adding the tag variable. Based on the tag-LDA model, they made real-time inferences about the likelihood of a particular tag being assigned to a new document, which is further used to generate recommended tags. In order to refine tags associated with images, Xu et al. (2009) proposed a regularized LDA (rLDA) which facilitates the topic modeling by exploiting both the statistical nature of tags and visual affinities of images in the corpus.

While these studies provide a good starting point for understanding the characteristics and uses of different social tagging systems, most of them have built their analysis on a static network or on a series of static snapshots of the evolving social network over various time periods. Few provide detailed analysis of the macro and micro features of the evolving tagging activities, focus on the evolution of social vocabularies or look for causes behind macro dynamic features coming from hidden patterns of individual taggers and resources. Cattuto, et al (2007, 2009) propose a dynamic vocabulary growth model that is modified and adopted in the first part of this paper: our main methods are inherited from Cattuto. In contrast to Cattuto's research, we applied the model to a comprehensive comparison of the evolving social vocabularies of the three most popular social tagging systems, Delicious, Flickr and YouTube, with broader data coverage over a more substantial time span. By using comparative analysis,

we found a relationship between the growth of global tags and the activity level of micro taggers. Our work also found the regularity of tag growth for the sub-groups of popular resources over physical time and the features of taggers' tagging activities over intrinsic time. We observed an exponential distribution of different sub-groups of resources in three tagging systems, including popular, less popular and non-popular resources. The applications of Latent Dirichlet Allocation (LDA) and adapted LDA models for social tagging systems proposed in previous studies focus on tags and neglect the dimension of taggers and the dynamics of topics over time. In this paper, we pay special attention to identifying reasons for the appearance of macro features of social vocabularies from individual resources. Finally, we propose a TTR-LDA model that extends LDA by incorporating all three elements of the social tagging systems, the tagger, resources and tags, in order to analyze the dynamic features of topics provided by these elements over time.

3 METHODOLOGY

3.1 *Data Collection*

We develop a tag crawler based on the Upper Tag Ontology (UTO) to harvest, integrate and store tagging data in RDF triples from Delicious, Flickr and YouTube. 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 (2009). There are two different parsers in the UTO crawler: one parses a page and searches for links that should be visited or filtered, while the other parses HTML code to retrieve data about tags in accord with the UTO. The working function of the crawler software in three tagging systems is described below (Ding et al., 2010):

In Delicious, the crawler began with the Delicious tag cloud at http://delicious.com/tag and visited every tag in the cloud. For TagA in the tag cloud, the crawler visited http://delicious.com/tag/tagA and parsed the HTML code to grab information about bookmarks, taggers and related tags. If a link is only bookmarked by one tagger, the tagging information of the tagger (taggerA) will be extracted from http://delicious.com/tag/tagA, otherwise the information will be extracted from http://delicious.com/url/idOfUrl, which also contains the information tagged by taggerA. For each bookmark having more than one tagger, the crawler then went to http://delicious.com/url/idOfUrl and crawled the history of the

bookmark, focusing on which users had tagged this bookmark on which date(s). After gathering data about all of the bookmarks on the first page for TagA, the crawler visited the second and subsequent pages for TagA, performing the same tasks.

For Flickr, the crawler started at the tag cloud at http://flickr.com/photos/tags and visited tags in the cloud. On each tag page (i.e. http://www.flickr.com/photos/tags/party/) information about related tags was collected. Each photo on the tag page (20 links per page) was visited (i.e. http://www.flickr.com/photos/ 25612622@N08/3063428352/) and information about the photograph, tags and tagger (one per photo) was extracted. The crawling process continued with http://www.flickr.com/photos/tags/party/?page=2. To avoid duplicate visits only links of the form http://www.flickr.com/photos/ taggerID/photoID/ were accepted.

For YouTube, the crawler started from the main page at http://youtube.com and visited every available video page (links starting with http://www.youtube.com/watch?v). On one video page it collected tagging data and visited the links pointing to other video pages. YouTube does not provide related tag data. In order to avoid visiting the same page more than once, the query parts of links were ignored (i.e. http://www.youtube.com/watch?v=X2IExa2A198 and http://www.youtube.com/watch?v=X2Iexa 2A198&watch_response lead to the same video).

When the crawler reached a Webpage that contained tag data, it sent information (including taggers who have created the tag, the resource link which the tag is used to describe, time when the bookmarking activity happened.) to Jena, where the information was stored in RDF format.

To store the tag data information, a Jena model (internal representation of an RDF graph) was created at program startup, after which the general UTO properties (has_source, has_object, has_comment, has_tag, has_date, has_tagger, has_related_tag and has_vote) were created. The information was added on the fly during the crawling process. If a certain configurable timeout was reached the model was written to the hard disk and the memory internal model was reset. The result was multiple plain text files containing RDFXml formatted triples.

In general, the crawler collects data from the HTML coding and populates the elements of UTO accordingly. For example, when the crawler reaches a Webpage that contains tag data, it

sends the information to Jena (http://jena.sourceforge.net/), which stores the data according to the UTO (Ding et al., 2010).

In November 2008, The UTO crawler was used to retrieve tagging data from Delicious, Flickr and YouTube. The crawler identifies objects, taggers, tags, dates, comments and votes. In total, the data retrieved contains approximately 3 million bookmarks, 0.6 million taggers and 15.7 million tags harvested from Delicious; 1.4 million photos, 0.07 million taggers and 17.7 million tags harvested from Flickr; and 1.4 million videos, 0.8 million taggers and 11.3 million tags harvested from YouTube.

3.2 *Data processing*

The crawled dataset covers Delicious from 2003 to 2008, Flickr from 2004 to 2008 and YouTube from 2005 to 2008. We use unified format {tagger, link, tag {tag 1, tag 2, tag 3… tag k}, time} to represent one post. A post is a tagging event in which one tagger tags one object with one or several tags. To further process the data, we delete data that existed before the system was established (for example, there are some tags in Delicious that appeared before 2003), posts with missing values (such as no tagger, no link, no tag or no date), and repeated annotation activities of taggers (where a tagger may bookmark the same link with the same tag more than once).

3.3 *Experimental Data*

After data processing, we obtain 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 1: 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

3.4 *Macro and Micro tag growth and micro taggers growth algorithms*

We build a dynamic tag growth model based on a time counting variable *tg* which is taken from the definition of intrinsic time (Cattuto, et al, 2007, 2009). Within our dataset, all posts

are sorted by date in ascending order, and the number of tags is initially set at 0. Each time one post is added, we count the number of tags in that post as m, and update tg as $tg=tg+m$. In previous studies related to the dynamics of social tagging systems (Golder & Huberman, 2006; Halpin et al., 2007), data are usually chopped according to time periods (days, months, and years), which has little to do with the actual period of tagging behaviors. For example, people may tag 100 bookmarks one day and none another day. By contrast, the dynamic tag growth we introduce here 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. Based on the above definition, we propose to use three main algorithms to evaluate the dynamic features of social tags, following Cattuto et al (2007, 2009).

Macro Tag Growth Method

The Macro Tag Growth Method (MaTG) calculates the evolution of tags at the macro level, measuring the global 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. The process is briefly described as follows: all the posts are sorted by their dates of creation from the earliest to the latest. The value of the $(tg_n, f(tg)_n)$ pair at the creation of the n^{th} post is obtained by increasing tg_{n-1} by the number of tags in the n^{th} post and increasing $f(tg)_{n-1}$ by the number of new tags in the n^{th} post; $(tg_{n-1}, f(tg)_{n-1})$ is the value pair at the creation of $n\text{-}1^{th}$ post.

Micro Tag Growth Method

The Micro Tag Growth Method (MiTG) measures the micro level of a social tagging system, analyzing the dynamic features of individual resources within it. We call these individual resources "target resources." MiTG is very similar to MaTGA except for a slight change, from selecting all the posts to selecting those associated with individual resources. For example, if the target resource is www.facebook.com, only posts that bookmark this resource are collected and analyzed.

Micro Taggers Growth Method

The Micro Taggers Growth Method (MiTaG) calculates the growth of taggers $U(tg)$ who bookmark a certain resource as the function of tg. The process is briefly described as follows: all the posts associated with an individual resource are selected and sorted by their dates of creation from the earliest to the latest. The value of the $(tg_n, U(tg)_n)$ pair at the creation of the

n^{th} post is obtained by increasing tg_{n-1} by the number of tags in the n^{th} post and increasing $U(tg)_{n-1}$ by the number of tagger in the n^{th} post (i.e., 1); (tg_{n-1}, $U(tg)_{n-1}$) is the value pair at the creation of $n\text{-}1^{th}$ post.

3.5 *TTR-LDA Model*

The algorithm introduced in Section 3.4 is mainly used to observe the cognitively regular patterns of tagging activities from the perspective of statistics. However, it is interesting to know how to make use of those cognitively regular patterns to optimize text mining for bookmarking systems. LDA (Latent Dirichlet Allocation) provides a solution to find semantic meanings from social tagging systems.

To analyze the dynamic features of topics provided by taggers, tags and resources over time, we propose a new model named Tagger Tag Resource LDA (TTR-LDA). This model extends LDA by incorporating all the three elements of social tagging system, the tagger, resources and tags.

LDA is generally based on the hypothesis that a person writing a document has several related sub-topics in mind. Tao address a topic, the author needs to pick a word with a certain probability of usage from the pool of words related to that topic, as well as other subjects included in the document. A whole document can then be represented as a mixture of different topics. When the author is one person, the chosen topics reflect his/her viewpoint and particular vocabulary. In the context of tagging systems where multiple users are annotating resources, the resulting topics reflect a collaborative shared view of the document and the tags related to the topics reflect a common or agreed-upon vocabulary. Since Blei et al. (2003) proposed the LDA model in 2003, it has been adopted by many researchers in different disciplines. This model can be used to analyze large quantities of documents, and to identify topics from those documents at a relatively high accuracy level. By using the LDA model, we can locate topics and their probability distribution in a set of documents, as well as the representative keywords and their probability distribution in each topic.

The proposed TTR-LDA model extends the LDA model by incorporating tags, taggers and resources (Tang, Jin, & Zhang, 2008). Our TTR-LDA model is described in Figure 1:

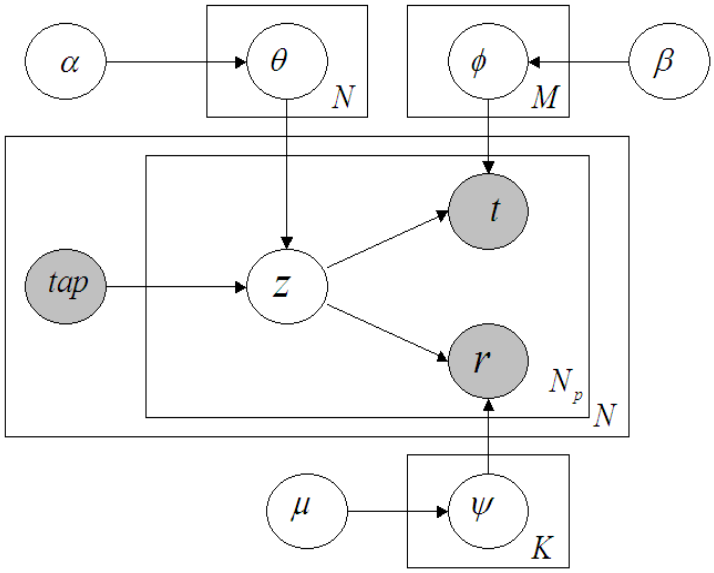

Figure 1：TTR-LDA Model

We assume that we have N posts (where one post can be seen as the activity that a tagger bookmarks tags for a certain resource), M distinct tags, K resources and T topics. The process works as follows:

1. Choose $\theta \sim Dir(\alpha)$, $\phi \sim Dir(\beta)$, $\psi \sim Dir(\mu)$;
2. For each post p:
3. For each tag t_{pi} in p:
4. For tagger ta_p, resource r_p in p, choose a topic z_{p_i} for t_{pi}, ta_p and r_p according to multinomial($\theta_{ta_p} \times \phi_{z_{pi}} \times \psi_{z_{pi}}$);

We can derivate the equation from the above process as follows:

$$P(z_{pi} \mid z_{-pi}, ta, t, r, \alpha, \beta, \mu) \propto \theta_{ta_p z_{pi}} \times \phi_{z_{pi} t_{pi}} \times \psi_{z_{pi} r_p}$$

$$\theta_{ta_p z_{pi}} = \frac{m^{-pi}_{ta_p z_{pi}} + \alpha}{\sum_{z'} m^{-pi}_{ta_p z'} + T\alpha}, \phi_{z_{pi} t_{pi}} = \frac{n^{-pi}_{z_{pi} t_{pi}} + \beta}{\sum_{t'} n^{-pi}_{z_{pi} t'} + M\beta}, \psi_{z_{pi} r_p} = \frac{n^{-pi}_{z_{pi} r_p} + \mu}{\sum_{r'} n^{-pi}_{z_{pi} r'} + K\mu},$$

Where $m^{-pi}_{ta_p z_{pi}}$ means the number of times topic z_{pi} is being assigned to tagger ta_p excluding the current one; $n^{-pi}_{z_{pi} t_{pi}}$ means the number of times tag t_{pi} is being assigned to topic

z_{pi} excluding the current one; $n^{-p}_{z_{pi}r_p}$ means the number of times resource r_p is being assigned to topic z_{pi} not including the current situation.

For the estimation of hyper parameters α, β, μ, we assigned different values for each hyper parameter and ran the TTR-LDA model to get the results. After several rounds of experiments, we found that different values of hyper parameters have little influence on the performance of the TTR-LDA model, consistent with Lu, Hu, Chen et al's results for Delicious (2009). Using the estimates provided in Tang, Jin, and Zhang (2008), we assign the hyper parameters as: α=50/K (where K is the number of topics), β=0.01 and μ=0.1.

4 RESULTS AND DISCUSSION

We compare the dynamic tag features of the three systems from the both macro and micro perspectives. The results of the macro level analysis of three tagging systems are discussed in Subsection 4.1.1 and the results of the micro analysis in Subsections 4.1.2, 4.1.3, 4.2.1 and 4.2.3.

4.1 *Comparison of macro dynamic feature in three tagging systems*

4.1.1 Comparison 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 the three applied tagging systems. The results are graphed into a log-log plot for *f(tg)* and *tg* are shown in Figures 2a, 2b and 2c. 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. The dashed line provides a linear approximation, and the small figures embedded in Figure 2a, 2b and 2c show the new vocabulary growth of $f(t)$ as function of physical time *t*.

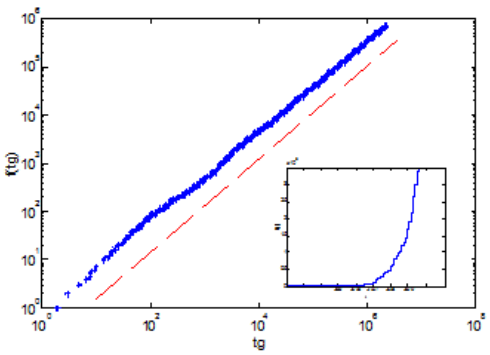

Figure 2a: Global Dynamic Growth of Delicious, where the dashed line closely meets the power-law with exponent $\gamma \approx 0.8040$.

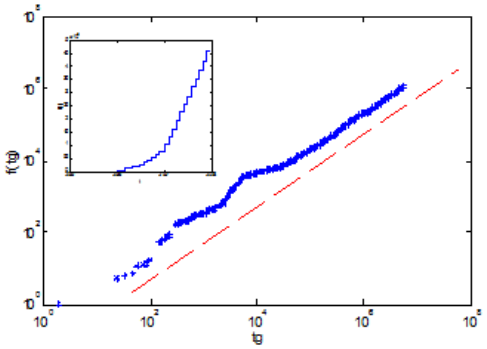

Figure 2b: Global Dynamic Growth of Flickr, where the dashed line closely meets the power-law with exponent $\gamma \approx 0.8039$.

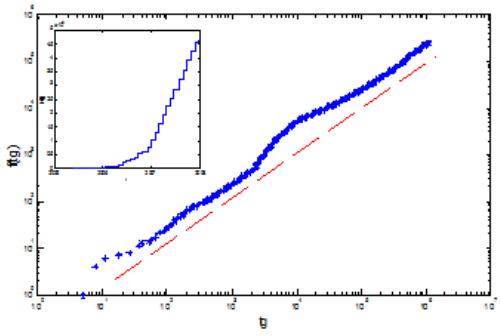

Figure 2c: Global Dynamic Growth of YouTube, where the dashed line closely meets the power-law with exponent $\gamma \approx 0.8580$.

In order to better observe the vocabulary growth of three tagging systems over physical time, we introduce time counter variable dt. We ranked all posts according to the time of their occurrences, and dt is the number of each post's rank of time. Then we analyzed the vocabulary growth of three tagging systems over dt, and found that the growth satisfies the Heaps' law, $f(dt) = K \times dt^{\beta}$ (Heaps, 1978). The results presenting the new vocabulary growth *f(dt)* as the function of time counter variable *dt* can be seen in Figure 3 below (we make log-log for both X axis and Y axis):

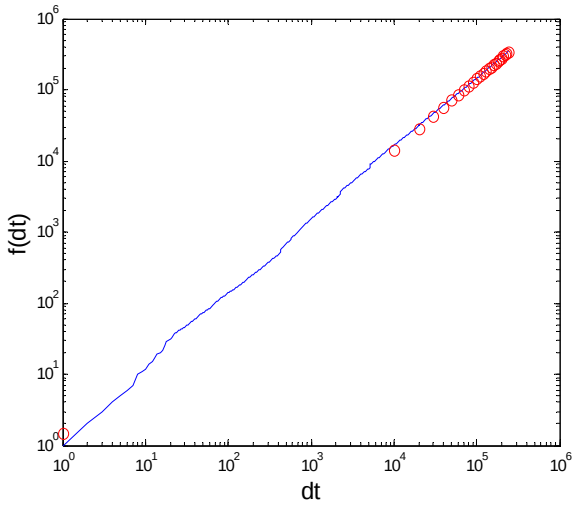

Figure 3a: Curve of f(dt) in log scale as the function of dt in log scale in Delicious

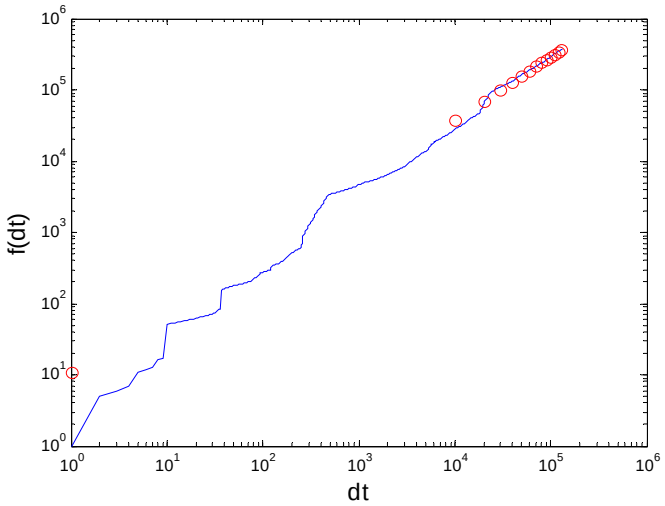

Figure 3b: Curve of f(dt) in log scale as the function of dt in log scale in Flickr

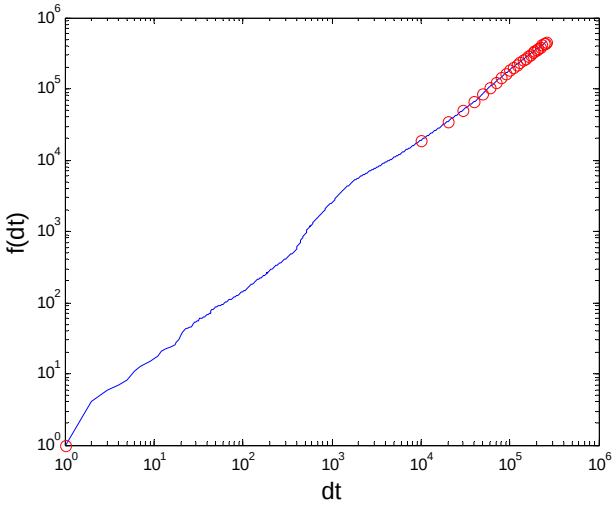

Figure 3c: Curve of f(dt) as the function of dt in log scale in Youtube

The coefficient and correlation coefficients of fitting function in three systems are listed in Table 2:

Table 2: The coefficient list of fitting function

	K	β	***Correlation coefficient***
Delicious	10.6684	0.9964	0.9906
Flickr	1.4739	0.8878	0.9842
Youtube	2.6713	0.9663	0.9917

As can be seen in Figure 3 and Table 2, Delicious has larger K and β than Flickr and Youtube, which means that taggers tend to create more new tags on this site than the other two during a fixed time period Δdt. That phenomenon may be attributed to the content of three systems and the increasing rate of taggers and resources. For example, compared with videos (Youtube) and photos (Flickr), taggers may use more distinct vocabulary to describe a webpage. The taggers and resources in Delicious are increasing more rapidly than those in Flickr and Youtube. Youtube also had larger K value and β value than Flickr. One important reason is that Youtube has more taggers than Flickr (about 10 times in our experimental data), so at the same time point, more tagging activities can occur in Youtube than in Flickr.

In addition, we found that there exist common features in three systems. For example, the growth rates of new vocabularies all exhibit decreasing rates over physical time t, and can not be influenced by other factors. We took the activities of the top 50,000 ranked taggers as an example (The data can be seen in Table 3 (Li, D., et al, 2010)). The tag growth from 2005-2007 in Delicious was rapid and the number of posts was relatively small (the total number was 231,199). In 2008, the rate of new tag growth decreased, while the total number of posts increased rapidly, with the total number of all posts reaching 1,108,782. This phenomenon indicates that the number of posts does not determine the rate of tag growth. This phenomenon can also be expressed as a cognitive process wherein new taggers tend to use existing tags to describe a certain resource rather than create new tags. The set of tags for a popular resource provides an accurate description of the content of that resource after a period of tag growth.

Table 3: Descriptive Statistics of Delicious data in the four time slices

	2005	2006	2007	2008
No. of posts	11,451	49,583	170,165	1,108,782
No. of resources	7,117	25,036	63,273	311,518
No. of taggers	3,616	12,053	28,823	48,688
No. of tags	10,014	31,493	78,661	283,188

The curve of vocabulary growth along intrinsic time with exponent less than 1, the phenomenon of which may be influenced by some factors, for example, the effect of the reuse/feedback mechanism enabled by the collective environment. We will use experiments to verify that in our future work. For each tagging network, the values of γ are different: $\gamma_{flicker} < \gamma_{delicious} < \gamma_{youtube}$. Different exponents of vocabulary growth reflects different natures of the three social tagging systems; different tagged objects (i.e., webpages, photos, and videos) lead to different purpose of collective endeavors and user motivations, which further poses impact on the growth and semantic structure of social tagging vocabulary. The exponent of YouTube is larger than Flickr and Delicious. Flickr and Delicious involve more individuals in the collective process through the social functions they provide, resulting in more variations from individual diversity. Videos on YouTube, however, tend to be tagged only by users who upload these videos, leading to a more semantically coherent vocabulary. A slower rate indicates a larger portion of tag reusing, which further implies a stronger collective feedback and higher-level consensus over time. A higher-level consensus consequently contributes to a more coherent and sense-making vocabulary. According to the results, the three social tagging systems under investigation present different rate of exponential growth, which indicates that the quality of social tagging vocabulary with regard to indexing and information retrieval is context-sensitive; i.e., different context (i.e., tagged object, system design, user motivations, etc.) will affect the quality of social tagging vocabulary. This outcome links to another discovery in Section 4.2.2, namely that, the mean exponent of the tagging activities of highly active taggers in these three systems are: $\overline{\gamma}_{flicker.tagger} > \overline{\gamma}_{delicious.tagger} > \overline{\gamma}_{youtube.tagger}$. This reveals that taggers in Flickr and Delicious tend to create new tags to tag resources, while YouTube taggers

tend to use existing tags to tag videos. This also helps explain why the YouTube tag vocabulary is more stable than that of Flickr and Delicious.

We also find that the value of γ is very similar in different social tagging systems, where it ranges from 0.8 to 0.9. From the perspective of linguistics, social tagging vocabulary is natural language created by human beings with tag as the unit. Therefore, in order to reflect the uniqueness of creation of tags, we compare the growth of unique tags along with the growth of the whole corpus in social tagging vocabulary with other kinds of corpora. In other systems, such as English corpora, γ ranges from 0.4 to 0.6 (Harman, 1995), and in Thai subset of WWW webpages, it is close to 0.5 (Sanguanpong, Warangrit, & Koht-arsa, 2000). The reasons for the difference in the value range of γ between social tagging systems and English corpora are (Cattuto, 2007): (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 (Veres, 2006).

4.1.2 Comparison of probability distribution of average "post length" in three tagging systems

Post length means the number of distinct tags in a post (Cattuto et al., 2007). The probability distribution of global post length in the three systems is shown in Figures 4a-c.

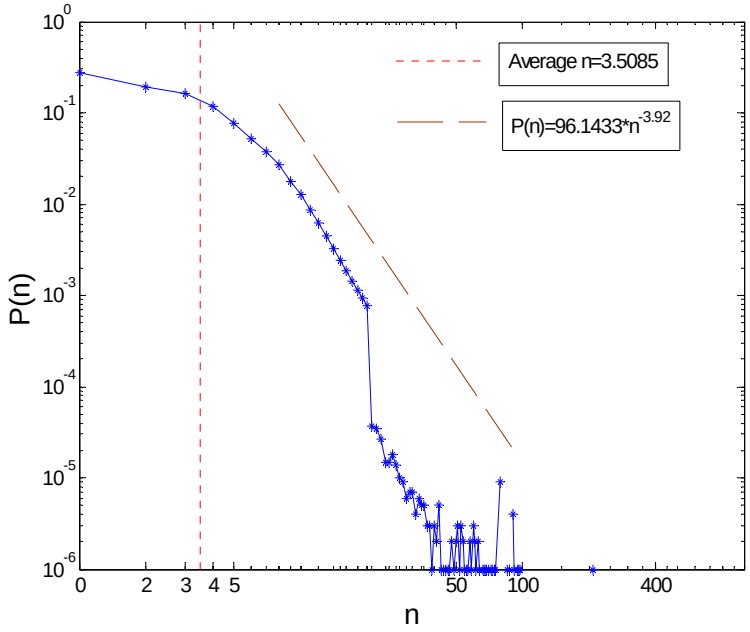

Figure 4a: Probability Distribution of average "Post Length" in Delicious.

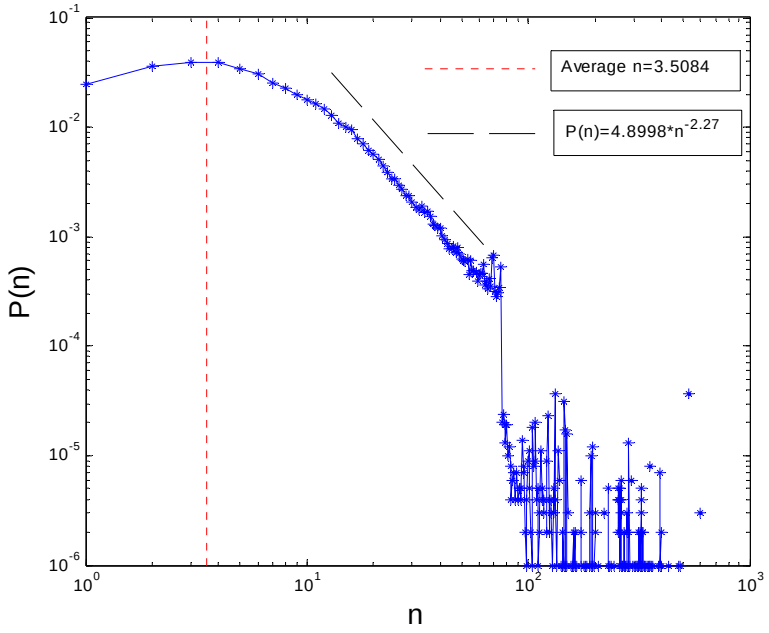

Figure 4b: Probability Distribution of average "Post Length" in Flickr.

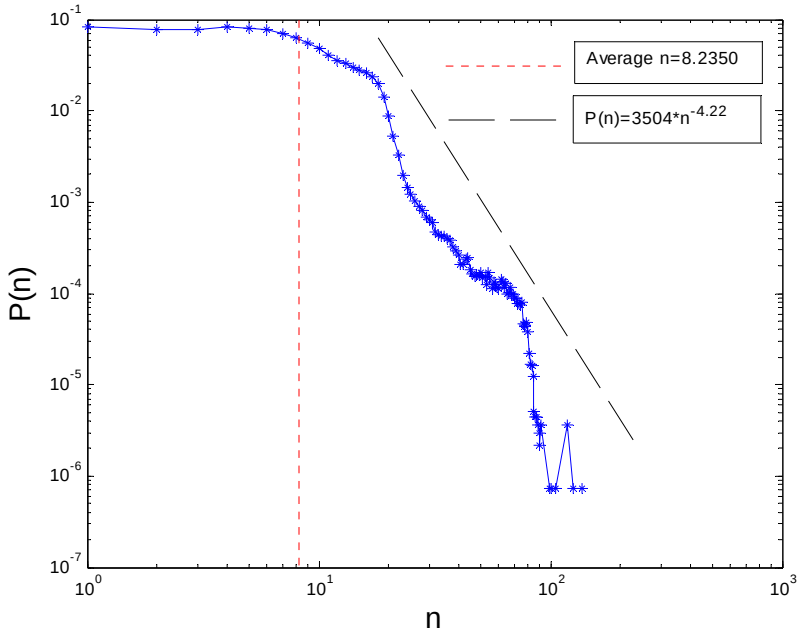

Figure 4c: Probability Distribution of average "Post Length" in YouTube.

The vertical dashed line represents the average "post length" of certain tagging systems and the gradient dashed line represents the exponents of the power-law tails of the curve. We found that the average post length of YouTube is the biggest at 8.2350, while the average post

lengths of Delicious and Flickr are similar at 3.5085 and 3.5084, respectively. They all display the steep decrease on their fat power-law tail. The decrease exponent of YouTube is the highest at –4.22, Delicious is second at –3.92, and Flickr is the lowest at –2.27. According to Figures 4a-c, we find that all three systems satisfy Zipf's distribution, as the exponents of their power-law tails are similar.

4.2 Comparison of micro dynamic features in three tagging systems

4.2.1 Comparision of micro tag growth in three tagging systems

Cattuto et al. (2007) have proved that the macro tag growth exponent is similar to the micro tag growth exponent of popular resources in Delicious (micro tag growth means the tag growth of a certain resource), captured here by using MiTGA. We select ten out of the 1,000 most popular resources, taking one after every 100th of top-ranked resources. We draw lines of tag growth as the function of *tg* for each resource (Figure 5).

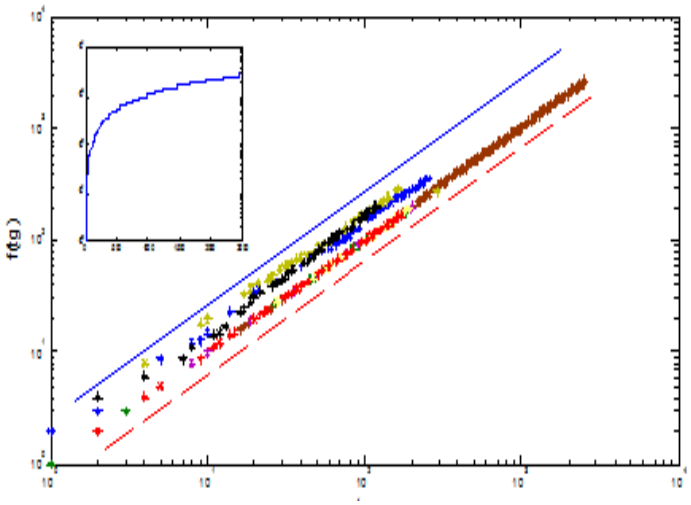

Figure 5: Tag growth for ten popular resources in Delicious (in log-log scale).

In Figure 5, the tag growth of all ten popular resources shows a sub-linear feature with parallel consistent growth after a period of time. This growth satisfies the sub-linear power-law distribution, where the resource exponents are between 0.5786 and 0.9245 (represented by the solid and dashed lines). Also, the slope of the micro tag growth of each resource decreases over with physical time *t* (shown in the small figure embedded in Figure 5), which means the rate of creating new tags becomes lower and will reach a fixed value after a period of time. The same analysis could not be conducted for Flickr and YouTube because the number of taggers who

tag popular resources is too small. This fact also reflects the differences found in the intrinsic nature of tagging behavior across the three social tagging systems. For example, in Flicker, there are restrictions of "self-tagging" and "permissive-tagging," while in YouTube, uploaders account for the dominant portion of taggers. In this case, even popular tags are tagged by a relatively small number of users.

4.2.2 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 $\gamma_{micro} = \log(f(tg_{\max}))/\log(tg_{\max})$ (Cattuto, et al, 2007). For our Delicious, Flickr, and YouTube datasets, we rank all the resources according to the number of taggers associated with them and select the top 1,000-ranked, 100,000-101,000-ranked and the lowest-ranked 1,000-ranked resources in each social tagging system. Those three groups of resources are considered to be popular, less-popular and non-popular within each system.

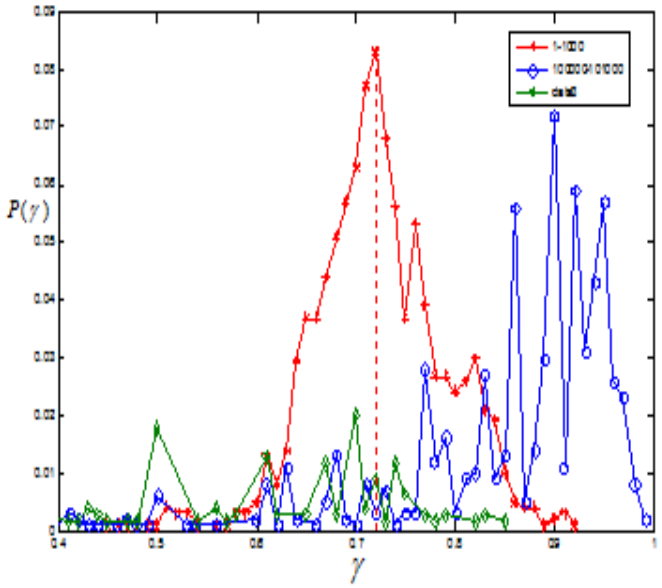

Figure 6a: Exponent probability distribution of groups of resources in Delicious.

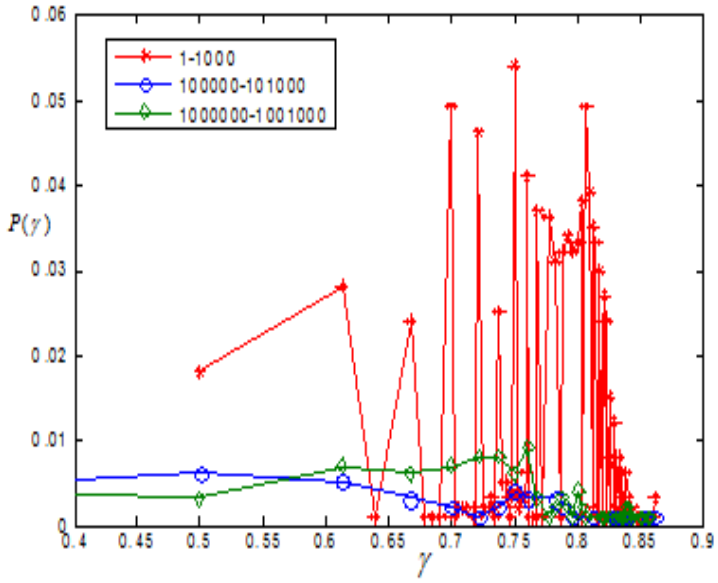

Figure 6b: Exponent probability distribution of groups of resources in Flickr.

In Figure 6a, the exponent of tag growth of the top 1,000-ranked resources at the final spot of tg in Delicious follows normal distribution with the mean value of 0.72, while less-popular and non-popular resources do not follow normal distribution. In Figure 6b, the top 1,000 ranked resources in Flickr have not yet reached the same distribution as top 1,000 ranked resources in Delicious. The explanation may be that Flickr imposes restrictions on tagging, allowing only "self-tagging" and "permissive-tagging" noted above. If the number of different taggers who tagged those popular resources can differ according to exogenous restrictions, this may result in non-normal distribution of the exponent probability distribution of Flickr's popular resources. As the number of taggers per resource is less than two in YouTube, and as majority of resources has similar exponent values, it is very difficult to calculate the probability distribution of its exponent.

In order to understand the normal distribution of tag growth exponents in the top-ranked Delicious resources, we select the top 5,000-ranked resources and use the same method to compute the exponent probability distribution of each resource. We calculate the timeline for each resource by subtracting its latest date of tagging from its earliest date of tagging and divide this timeline into four stages. For each stage, we compute the tag growth exponent probability distribution of the selected 5,000 resources (Figure 7).

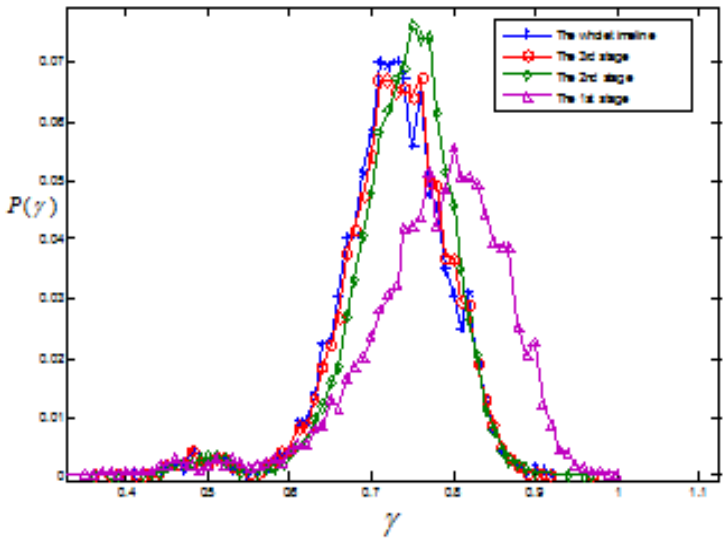

Figure 7: Tag growth exponent probability distribution of resource groups in different developmental stages.

Figure 7 provides insight into how popular resources are formed. The values of skewness and kurtosis of exponents distributions for each time period can be seen in Table 4.

Table 4: Skewness and Kurtosis for each time period

	The 1st stage	The 2nd stage	The 3rd stage	The 4th stage
Skewness	-1.2136	-0.9182	-0.9110	-0.7877
Kurtosis	-0.5337	-0.2272	-0.1246	0.0913

As can be seen in Table 4, all the Skewness values are smaller than 0, which means all the distributions have a fat tail on the left, the absolute values of kurtosis and Skewness become smaller over physical time. The value of the exponent approaches 0.7. Skewness can be described as the "measure of the asymmetry" of the normal distribution while kurtosis can be seen as the "peakedness" of a normal distribution. According to the changes of kurtosis, skewness and mean value over physical time, we can describe the exponent distribution as: its "peakedness" will approach to the shape of normal distributions; its "height" will become higher, while its crest moves to the left over the physical time. It indicates that the asymmetry of exponent distribution is significant.

After further testing on all the Delicious resources, we find that the average value of γ_{micro} of each resource is around 0.72 (with a standard error of 0.02). We also find that the proportion of the resources whose average γ_{micro} is between 0.7 and 0.74 is around 3.7%. For all the resources in Delicious, referring to Cattuto's research (2007), we found that when the

resources become popular enough, their exponents of vocabulary growth have a higher probability of approaching 0.72 (Section 4.1.3, last paragraph). Thus, we believe that 0.72 may be an intrinsic feature of the set of popular resources in social tagging.

4.2.3 Comparision of micro tagger growth in Deliclious

We rank all the resources by their total number of distinct tags in each of the three tagging systems, and then select ten popular resources from the top 1,000 resources using the procedure explained in Section 4.1.3. We randomly choose one resource from the top 100 resources and then select one after every 100^{th} resource based on the first one. For each resource, we use the MiTaGA algorithm to compute the tagger growth as the function of *tg*. The dynamic tagger growth on Delicious is shown in Figure 8.

We can see that the distributions fit the power law distribution well. We observe (Figure 8) that although there is noise at the early stage of tagger growth for ten popular resources, after a period of time (when they become popular enough), the curve track of all the resources tends to become unified. Their exponents $\gamma_{U(tg)}$ converge to 0.3190, suggesting that each tagger tends to provide the same number of tags for those popular resources when they become popular enough. This consistency may be due to the various social functions that support the reaching of consensus among individual taggers. We use the same method to select five resources from less-popular resources and non-popular resources respectively. The curves of tagger growth for each resource are shown in Figures 9 and 10. In Figures 9, for the five less-popular resources, curves show a clear trend of convergence. In Figure 10, for the five non-popular resources, curves display different slopes and variations, suggesting that individual taggers assigned tags to resources based on their own contexts, interests, and concerns.

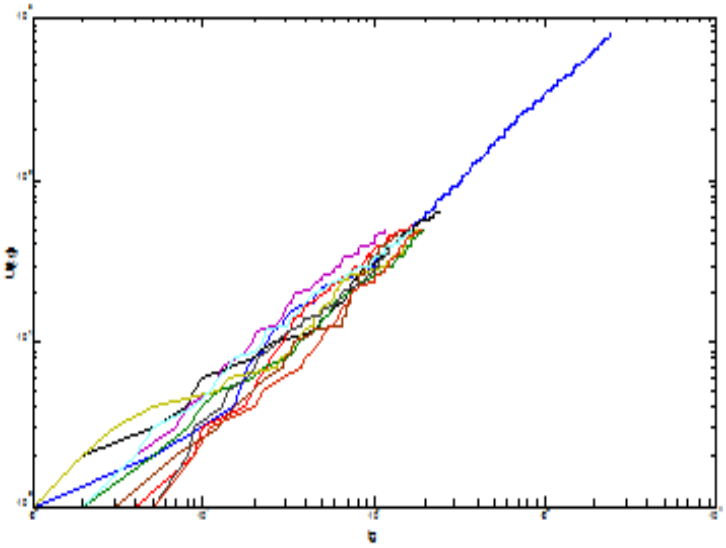

Figure 8: Micro taggers growth of ten popular resources as the function of *tg* in Delicious (in log-log scale).

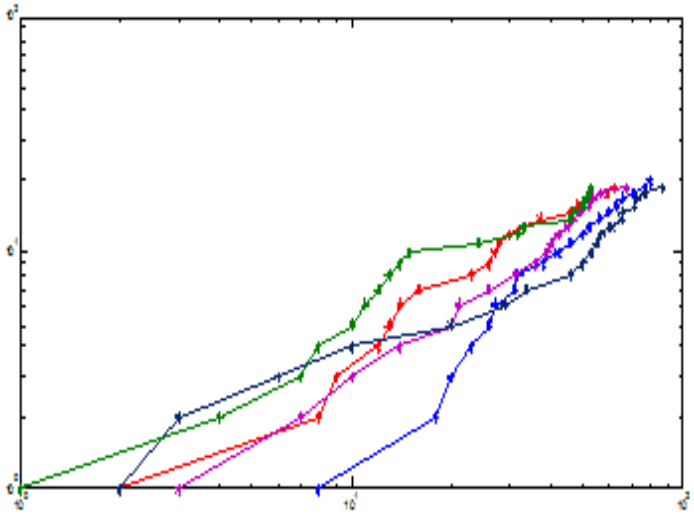

Figure 9: Micro tagger growth of five less-popular resources as the function of *tg* in Delicious (in log-log scale).

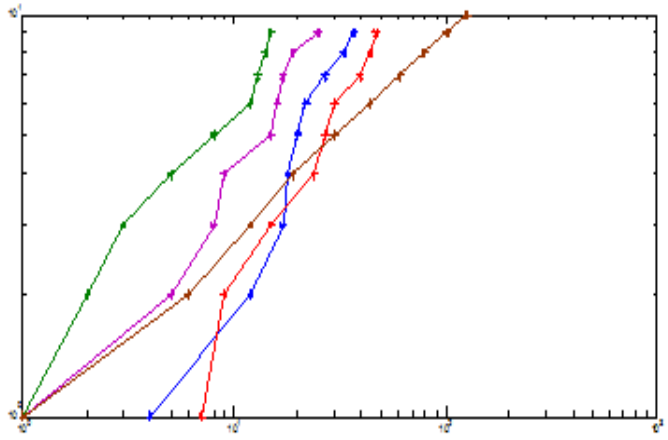

Figure 10: Micro tagger growth of five non-popular resources as the function of *tg* in Delicious (in log scale).

As shown in Figures 8, 9, and 10, the inverse of the slope denotes, at that point, the number of tags which an additional tagger created to tag the resource. In Figure 9, when the resources become popular enough, taggers tend to use the same number of tags to tag the ten selected popular resources. In Figure 10, the slope of tagger growth of the selected ten less-popular resources shows a clear convergence, while in Figure 10, tagger growth rates of different non-popular resources vary from each other. Different levels of convergence shown in Figures 8, 9, and 10 correspond with the popularity levels of resources respectively. The more popular the resources are the higher level the convergence of tagger growth is. This phenomenon is reasonable, because higher popularity indicates a deeper impact from tag reuse/feedback mechanism.

4.2.4 Comparison of tag growth exponent probability distribution for highly active taggers in three tagging systems

We select the top 1,000-ranked taggers from each system, divide them into three groups and analyze their exponent of probability distribution. We find that all groups display a Gaussian distribution with different mean values $\overline{\gamma}_{tagger} : \overline{\gamma}_{flicker.tagger} > \overline{\gamma}_{delicious.tagger} > \overline{\gamma}_{youtube.tagger}$. The mean values of these systems reflect the tagging activity of highly active taggers: high value means that taggers tend to use new tags to tag new resources and low value means that taggers tend to use their existing tags to tag new resources. This confirms the outcome shown in Figure 2; since videos on YouTube are mostly tagged by users who upload them, the vocabulary of YouTube shows greater semantic coherence around content. Contrary to the

results presented in Figure 6, the tagger activities in these three tagging systems demonstrate normal distributions, while probability distribution of tag growth exponents in Flickr shows non-normal distribution.

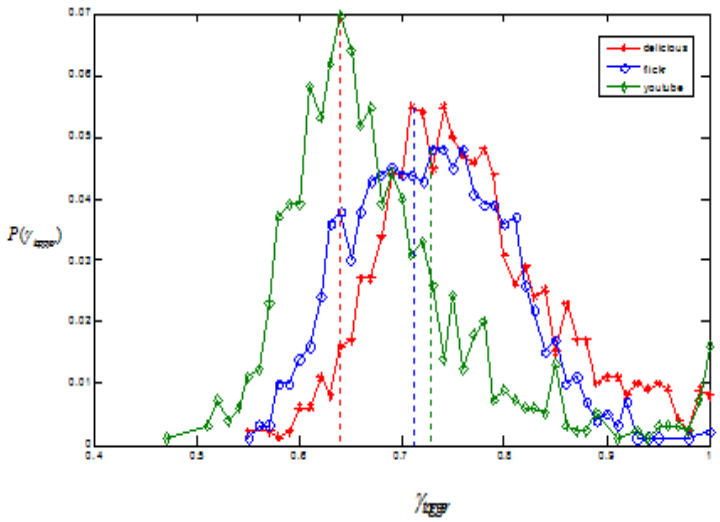

Figure 11: Probability distribution of different groups of taggers' activities in Delicious, Flickr and YouTube.

4.3 *The Dynamic topic modeling of Tagger-Resources-Tag Features of Delicious*

4.3.1 TTR-LDA for popular, less popular and non-popular resources

Based on the analysis above, we can see that about 95% of the top 1,000 ranked resources in Delicious have reached a stable status in which their increasing exponents are between 0.6 and 0.9, but that there is no similar phenomenon in YouTube and Flickr. In this section, we use our new TTR-LDA model to analyze the dynamic features of topics from these stable status associated with the top 1,000 ranked resources, 1,000 less-popular and 1,000 non-popular resources on Delicious to observe its topic distribution (see section 4.1.3). Our experiment on the top-ranked 1,000 resources includes 1,000 different links, 44,204 different taggers and 26,065 distinct tags. First, we randomly select a tagger and all his/her tags for all the resources he/she has bookmarked in the whole data set (2005-2008) as testing data (1,000 resources), whereas the rest are seen as training data (3,018 resources). We assign the number of topics as 1-400 respectively and input the training data into the TTR-LDA model to get the results. Next, we use test data to compute perplexity for each result (Michal et al., 2004). Perplexity shows the performance of a statistical model: the lower the perplexity value is, the better a model fits

the actual distribution. The perplexity value under different numbers of topics can be seen in Figure 12.

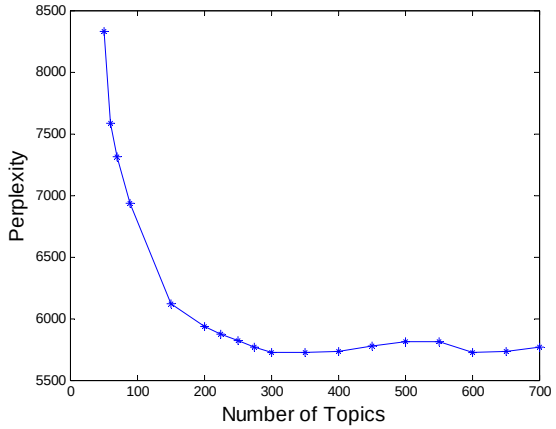

Figure 12: Perplexity value of different topics.

Figure 12 shows that the curve first gets the lowest value at the point around 300. After that, the curve displays that perplexity increases slowly with fluctuations. According to the previous studies (Blei, 2003; Michal, 2004), perplexity will not change significantly after it reaches a certain value, therefore, the number of topics for the whole data set was assigned as 300. We select the top 50 of these 300 topics and draw their probability distribution across resources (Figure 13).

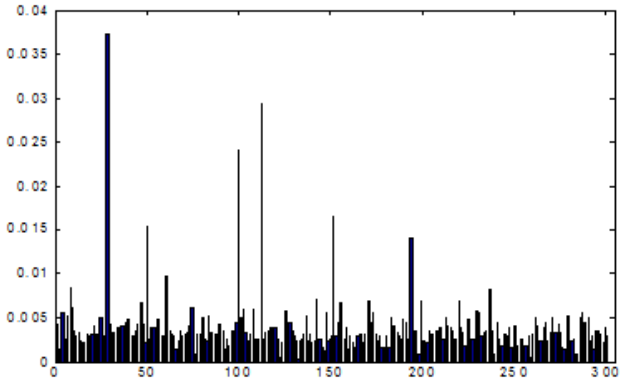

Figure 13: Probability distribution of 300 topics in 2008.

We select the eight most popular topics from all 300 topics, identify the five most representative tags (those with the highest probability) and taggers for each topic and summarize all the information in Table 5.

Table 5: The tagger, link information for top eight topics in the top 1,000-ranked resources.

Topic	Tpoic29	Topic112	Topic100
tag	frank%2Fgerard 0.061956 Bandom 0.045211 Bandslash 0.042935 Fic 0.034644 Slash 0.016761 bob%2Ffrank 0.011559 mychemicalromance 0.010909 bob%2Ffrank%2Fjamia 0.009445 rps 0.007495 pete%2Fmikey 0.007169	sga 0.075577 mckay%2Fsheppard 0.049748 fic 0.029703 eureka 0.025077 slash 0.024691 crossover 0.022764 earthside 0.019101 humor 0.014475 john%2Frodney 0.013897 fanfic 0.012355	rps 0.034775 jared%2Fjensen 0.029529 slash 0.023409 spn 0.020786 sam%2Fdean 0.018601 supernatural 0.017945 first-time 0.015759 fic 0.015322 schmoop 0.014885 rating%3Anc-17 0.004830
link	Pearl-o.livejournal.com/1000307.html impertinence.livejournal.com/279588.html battleofhydaspe.livejournal.com/11472.html community.livejournal.com/inlipstick/9690.html #cutid1 mxtape.livejournal.com/59808.html	http://amific.livejournal.com/9577.html http://trinityofone.livejournal.com/25344.html http://community.livejournal.com/sga_flashfic/151595.html http://kashmir1.livejournal.com/788568.html http://www.intimations.org/fanfic/stargate/contradiction.html	http://gekizetsu.net/sn/twokinds.htm http://www.intimations.org/fanfic/supernatural/Kings%20and%20Queens%20and%20Jokers%20Too.html http://www.gekizetsu.net/sn/90proof.htm http://stele3.insanejournal.com/129835.html http://nutkin.livejournal.com/24724.html
Topic	**Topic153**	**Topic52**	**Topic194**
tag	apache 0.016009 webdev 0.008028 javascript 0.006619 ajax 0.006150 web2.0 0.005680 framework 0.005211 ui 0.005211 module 0.003803 plugin 0.003590 hibernate 0.002683	opensource 0.013597 php 0.010085 framework 0.007074 web2.0 0.004566 opac 0.004566 osx 0.004064 webdevelopment 0.003562 freeware 0.003061 lunix 0.003061 mac 0.002559	design 0.010270 inspiration 0.009538 art 0.007343 portfolio 0.005635 webdesign 0.005635 graphics 0.005391 illustration 0.005147 graphic 0.005147 gallery 0.004903 artist 0.002708
link	http://www.raibledesigns.com/tomcat/boot-howto.html http://ws.apache.org/axis2/ http://www.opencalais.com/ http://www.webappers.com/2008/11/05/best-cheat-sheets-for-web-developers/ http://snook.ca/archives/javascript/jquery-bg-image-animations/	http://www.sxc.hu/index.phtml http://sourceforge.net/ http://www.opengoo.org/ http://www.virtualbox.org/ http://virtuemart.net/ http://www.crystalspace3d.org/main/Main_Page http://code.google.com/p/waf/	http://www.sxc.hu/index.phtml http://quality-and-free-vector-object-sets-to-beautify-your-designs/ http://profile.myspace.com/index.cfm?fuseaction=user.viewprofile&amp http://www.istockphoto.com/index.php http://www.sxc.hu/index.phtml
Topic	**Topic61**	**Topic236**	
tag	sysadmin 0.009815 linux 0.008592 authentication 0.008478 encryption 0.008118 unix 0.007123 syndication 0.003213 freeware 0.003207 security 0.002962 tcp 0.002496 hacks 0.002368	socialmedia 0.003677 web2.0 0.003677 socialnetworking 0.003416 blog 0.002373 community 0.002373 trends 0.002373 culture 0.002112 communication 0.002112 business 0.002112 socialnetworks 0.001852	
link	http://www.virtualbox.org/ http://openid.net/developers/specs/ http://www.scottklarr.com/topic/115/linux-unix-cheat-sheets-the-ultimate-collection/ http://www.personalfirewall.comodo.com/ http://datacent.com/hard_drive_sounds.php	http://www.spiegel.de/netzwelt/web/0 http://www.time.com/time/specials/2007/article/0 http://www.time.com/time/specials/2007/0 http://www.diigo.com/ http://www.facebook.com/	

From Table 5, we see that the top three topics are mainly about writers and works of fiction, that topic 194 is related to art and gallery activities, and that the other topics are relevant to

programming and computer science. Although the top three topics are all related to fiction, they emphasize different aspects: the most popular topic is involves "bandslash" fiction (a subgenre of fan fiction, bandslash fiction refers to the romantic or sexual pairing of same-sex bandmates), the second topic concerns supernatural fiction and the third concerns Stargate: Atlantis Fanfiction (SGA)[1]. The TTR-LDA model can reveal the latent semantic structure of those tags, where a similar phenomenon can be observed in other topics: topic 152 is mainly about Web development technology, topic 51 is about opensources software, topic 61 is about freeware and security, and topic 236 is about social networking.

To make a clear observation of topic distribution for all reources in Del.icio.us, we also select 1,000 less-popular resources (the 100,000-101,000-ranked resource in the whole dataset), and 1,000 non-popular resources (the lowest-ranked 1,000-ranked resources in the whole dataset) and use the TTR-LDA model to compute their topic distribution respectively. We assign the number of topics as 100. The top five-ranked topics and their representive tags for the 1,000 less-popular and the 1,000 non-popular resources, as listed in Table 6:

[1] Stargate Atlantis (often abbreviated as SGA) is a Canadian-American science fiction television series and part of Metro-Goldwyn-Mayer Inc. Stargate franchise.

Table 6: Tag information for top four-ranked topics in 1,000 less-popular and 1,000 non-popular resources.

	Topic 20	Topic 67	Topic 60	Topic 95
Less-popular	muscle 0.013031 bodybuilding 0.011740 way 0.010450 lifting 0.007870 routines 0.006580 bulking 0.005290 musclehow 0.005290 weight 0.003999 musclemuscle 0.003999 muscleweight 0.003999 exercise 0.003999	bandom 0.013987 bandslash 0.009832 fic 0.009832 au 0.009832 patd 0.008448 as3 0.007063 picspam 0.005678 flash 0.005678 actionscript 0.005678 nc-17 0.005678 fob 0.005678	microsoft 0.012980 advertising 0.011553 news 0.011553 youtube 0.010127 politics 0.008701 video 0.007274 funny 0.005848 movie 0.005848 business 0.005848 tv 0.004422 journalism 0.004422	politics 0.027557 obama 0.026115 media 0.013129 election 0.010244 blogs 0.004473 article 0.004473 gaming 0.004473 economics 0.003030 economy 0.003030 stocks 0.001587 book 0.001587
	Topic 84	Topics 88	Topic 58	Topic 99
Non-popular	sap 0.040582 sdn 0.031289 webdynpro 0.012701 interactive 0.009603 adobe 0.009603 java 0.006506 ftp 0.006506 flash 0.006506 community 0.003408 brian 0.003408 properties 0.003408 portal 0.003408	charity 0.012781 donate 0.012781 sponsor 0.006546 aids 0.006546 hiv 0.006546 hiv%2Faids 0.006546 lifebeat 0.006546 ushahidi 0.003429 army 0.003429 advertising 0.003429 hermsan 0.003429 training 0.003429	sap 0.022771 portal 0.013149 myaccount 0.003528 .net 0.003528 connect 0.003528 businesswarehouse 0.003528 ticket 0.003528 expire 0.003528 singlesignon 0.003528 banking 0.003528 mobilebanking 0.003528	bodybuilding 0.006801 training 0.006801 sports 0.006801 body 0.006801 supplements 0.006801 weightlifting 0.006801 exercises 0.003562 security 0.003562 muscles 0.003562 nutrition 0.003562

From Table 6, we can see that our TTR-LDA model can detect specific topics for less-popular and non-popular resources, indicating the effectiveness of this model. For example, topic 20 for less-popular resources and topic 99 for non-popular resources are obviously about bodybuilding. Topic 67 is very similar to topic 29 of popular resources (see Table 6), reflecting that some popular and non-popular resources are both related to bandslash fiction. Moreover, other topics identified in less-popular and non-popular resources are quite different from those identified in popular resources, suggesting that the TTR-LDA model exposes peripheral areas of interests for taggers and communities on Delicious.

4.3.2 Dynamic TTR-LDA model for observing topic evolution

In this section, we are mainly interested in two issues: first, for popular resources, how do topics evolve over time? Second, can TTR-LDA model get more accurate semantic information from increasing tagging activities? For the first issue, we use Dynamic TTR-LDA

model to observe topic evolution of popular resources in Delicious over time; for the second issue, we performed two experiments to verify our assumption.

In order to observe the dynamic semantic features of social tagging, and to see how topics evolve, we use the TTR-LDA model to obtain probability distributions of topics for taggers, tags and resources at the end of each year from 2005-2008. The number of topics is assigned to 300, which has been verified eariler in the section. In order to determine whether these topics in different time periods reflect the same topic over time, we use Pearson similarity measure to compute the similiarity bewteen two topics from different years (Ahlgren, Jarneving, & Rousseau, 2003). For example, we have topic *TA1* from year *A* and we would like to know which topic *TBi* in year *B* is the same as *TA1*, where the representative tags in topic *TA1* are $\{(t_{a11},\omega_{a11}),(t_{a12},\omega_{a12}),(t_{a13},\omega_{a13})...(t_{a1m},\omega_{a1m})\}$ and in *TBi* are $\{(t_{bi1},\omega_{bi1}),(t_{bi2},\omega_{bi2}),(t_{bi3},\omega_{bi3}),...(t_{bin},\omega_{bin})\}$ respectively, t_{xij} means the *jth* tag in topic *i* in year *x*, ω_{xij} means the probability of that tag in topic *j* in year *x*. For each topic *TBi*, we find the tag intersection *Vi* between *TBi* and *TA1* and then use the Pearson similarity expression to compute the similarity *SVi* between every *TBi* and *TA1*. The *TBi* that has the largest *SVi* value can be considered to be the same topic as *TA1*. A heatmap is drawn to show the similarity matrix between topics in 2007 and 2008.

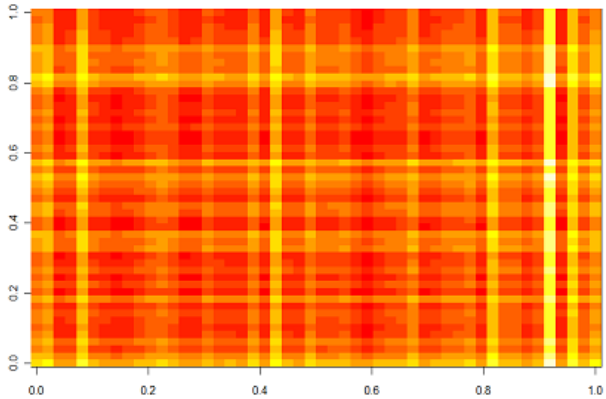

Figure 14: Heat map of 50 of 300 topics between 2007 (Y axis) and 2008 (X axis).

In Figure 14, the color of each cell is proportional to the similarity between the two topics that are correspodent to the cell from the X axis and Y axis respectively: the darker the color is, the higher the similarity value, and the lighter the color is, the lower the similarity. We can see

that many topics in 2008 are highly similar to topics in 2007 (the dark area), indicating that most topics share some tags with other topics. We also easily see that the 46th topic in 2008 shows a considerably low similarity to all the topics in 2007 (the column with nearly all cells in light grey). We find that the topic is associated with tags like mobile, iphone and PDA. The tag "iphone" presents a very low probability in 2007 and a high probability in 2008 across topics (it has the highest probablity in the 46th topic), suggesting a change of interest among taggers over time, with old topics vanishing and new topics emerging. When further checking iphone in Google Trends (http://www.google.com/trends?q=iphone), we find that the search volume index of iphone first appears in late 2004 but remains near 0 in 2005 and 2006; it has a first peak in early 2007 and then starts to grow, and in 2008 arrives at two other peaks. This increase in the search volume index conforms well to our results.

The dynamic TTR-LDA model can show specifically which topic is emerging and which topic is vanishing, from which the quick adaptability of social tagging vocabulary to users' interest change and external effect is shown. For example, "iphone" emerged as a topic shortly after it was introduced by Apple Inc. Compared with the inflexibility of controlled vocabulary, dynamic adaptability is an unique and effective feature of folksonomy.

We further caculate the similarity between topics in 2005 and 2008, between those in 2006 and 2008 as well as those in 2007and 2008, and identify the topics in each time period that have the highest similarity (the same topics) with the top five topics in 2008. Figure 15 shows the probability changes of the top five topics over the time. A higher probability means that the topic is more popular at that time period. We find that the topics about fiction (the top three-ranked topics in Figure 13) are the most popular topics in 2005, 2007, and 2008, but get a relatively low popularity during 2006. All the top three-ranked topics experience a similar change in the level of popularity over time. For example, Topic 4 and Topic 5 are related to computer science, and we can see that although their popularity levels fluctuate during 2005, their popularity tends to be closely correlated afterwards.

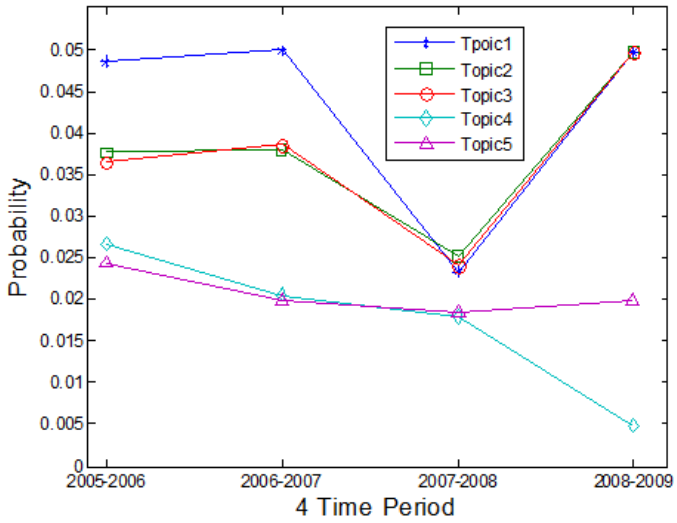

Figure 15: Probability distribution of the top 5 topics in four time periods.

From Table 6, we see that not only has the probability of topics varied across years, but also that the content of topics has changed frequently over time. We can observe, however, that there are no major content changes in the top five topics in 2007 and 2008.

Table 7: Changes of topic content over time.

Topic	Topic1	Topic2	Topic3	Topic4	Topic5
	bandom,	fic	sga	javascript	opensource
	bandslash	spn	mckay%2Fsheppard	webdev	linux
2008	fic	supernatural	fanfic	ajax	java
	slash	fanfic	fic	api	python
	rps	fanfiction	fandom%3Asga	plugin	oss
	bandslash	Fandom%3Asupernatural	mckay%2Fsheppard	webdesign	opensource
	bandom	sam%2Fdean	sga	css	plugin
2007	fob	spn	mcshep	javascript	cms
	falloutboy	supernatural	slash	webdev	php
	mcr	pairing%3Asam%2Fdean	fic	ajax	python
	fanfic	fandom%3Asga	fic	programming	python
	gen	angst	smallbiz	opensource	linux
2006	fic	author%3Ablueraccoon	programming.games	algorithms	parsing
	crack	fanfiction	silvertone	tools	ontology
	slash	human-aliens	mckay%2Fsheppard	plugins	mathematics
	articles	articles	articles	webdesign	.net
	literature	literature	literature	j2ee	csharp
2005	neilgaiman	neilgaiman	neilgaiman	java	opensource
	comics	comics	comics	myspace	cyberspies
	creativity	creativity	creativity	cyberspies	espionage

Compared with Figure 15, Table 7 shows the probability value of each topic displays a relatively smooth transition from 2007 to 2008. When we carry out a similar experiment for other topics, we find that if the content of the topics does not change much, the degree of popularity for that topic becomes stable for that period of time This may be explained by the evolution of the content of a topic into a relative stable stage within certain groups of taggers interested in that topic. We also find that a topic may evolve into different branches over time. For example, the top three topics belong to the same topic (fiction) during 2005, mainly about articles, literature and authors (as can be seen in Table 7). After three years of evolution, the topic has been divided into three new topics with new representative tags, and has entered a relatively stable status. We consider these three new topics as mature, in that they are associated with a stable set of tags used to describe themselves, and their popularity level does not change significantly over 2007 -- 2008 (Figure 15).

Additionally, we can build up an interest model for each tagger in social tagging systems by using TTR-LDA. Here we randomly select a tagger from the 1,000 most active taggers, and find the probability distribution of his/her interests over the 300 topics of that tagger at the end of 2008.

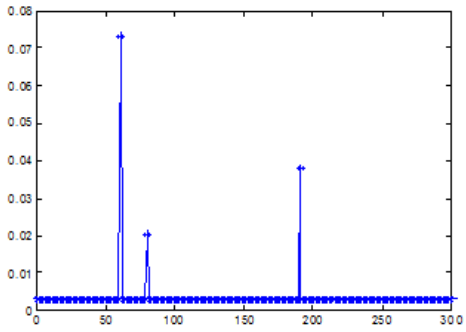

Figure 16: Tagger interest model over 300 topics.

As can be seen in Figure 16, the selected tagger is interested in topic 61 (online videos and movies), topic 80 (mashup, Web) and topic 191 (online music).

We used symmetric Kullback–Leibler (sKL) divergence (Rosen-zvi, M., Griffiths, T., 2004) to analyze the similarity between different resources pairs from the topic level and used the dynamic mechanism to observe their statistical features over time. We found that with

increasing bookmarking activity, more and more resource pairs exhibit apparent similarity from the topic level. The experiment results can be seen in Figure 17:

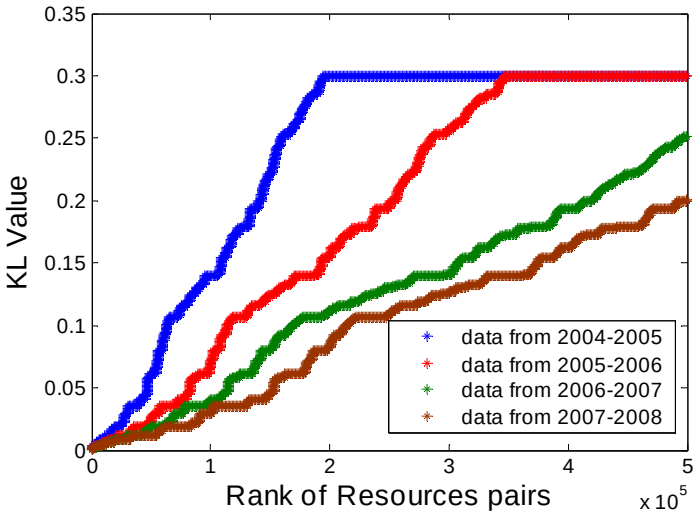

Figure 17: The sKL Divergence of resources pairs as the function of their ranks

As seen in Figure 17, the average sKL Divergence of resource pairs tends to become smaller and smaller over time. For example, for resource pairs with the same rank, the resource pairs during 2007-2008 have a lower sKL value than the resource pairs from other time periods. This suggests that more and more similar resource pairs are discovered with increased tagger activity. This phenomenon discloses the cognitive process of a tagger's tagging behavior.

In order to discover the differences between TTR-LDA and traditional methods such as TF-IDF for finding highly similar resources, we also counted the number of common tags for each resource pair. Different from traditional methods, which mainly focus on the common tags of two resources, TTR-LDA uses resources' topic distributions to compute their similarity. We found that traditional methods can find resources pairs with similar content in most cases, but they could not find highly similar resources pairs with low common tags. There are resource pairs,with highly related content, but the small number of common tags in our dataset. Most of them share few representative common tags, but their contents are highly relevant. In order to better illustrate our findings, we selected 10 representative resources pairs from the top 1,000 most popular resources, listed their sKL and number of co-occurrence tags in Table 7. The top 5 rows are resources pairs with low sKL and low common tags, while the bottom 5 rows are resources pairs with high sKL and high common tags.

Table 8: representative resources pairs with low KL divergence (top 5 rows) and high sKL divergence (bottom 5 rows)

Representative Resources pairs	*Number of co-occurrence tags*	sKL *divergence*
http://www.independent.co.uk/news/world/middle-east/our-reign-of-terror-by-the-israeli-army-811769.html *http://www.ynetnews.com/articles/0*	*6*	*0.000116*
http://www.w3.org/html/wg/html5/diff/ *http://deseloper.org/read/2008/04/a-simple-modal/*	*0*	*0.000325*
http://funkatron.com/site/comments/google-app-engine-from-a-php-developers-perspective/ *http://www.sitepen.com/blog/2008/06/05/easy-repeatable-buildingdeployment-of-pythondojo-projects/*	*4*	*0.000104*
http://www.oreillynet.com/ruby/blog/2008/09/inspect_sql.html *http://fuglyatblogging.wordpress.com/2008/10/*	*3*	*0.000053*
http://nyc.everyblock.com/ *http://streetclash.blogspot.com/*	*2*	*0.000388*
http://www.time.com/time/world/article/0 *http://www.foxnews.com/story/0*	*42*	*3.1687*
http://www.foxnews.com/story/0 *http://www.news.com.au/story/0*	*38*	*5.8595*
http://lifehacker.com/ *http://www.pcmag.com/article2/0*	*36*	*3.7546*
http://www.mediapost.com/publications/?fa=Articles.san&amp *http://www.spiegel.de/netzwelt/web/0*	*40*	*4.8996*

Lower sKL divergence means that the resource pairs have higher similarity at the topic level. We found from Table 8 that the resource pairs with low sKL divergence always have a low number of tag co-occurrence. Those resource pairs were judged as dissimilar according to traditional similarity methods, but when we checked their contents, we found that they have a high similarity from the topic level. Taking the first pair as an example, the first resource has 63 distinct tags while the second resource has 69 distinct tags. They only have 6 tags in common, but their contents are both about the military and politics. The same trend can be discovered in other resource pairs. In the second pair, the key tags of the first resource are about web development while the key tags for the second resource are about AJAX and java. In the fifth resource pair, the first resource is an overview of the blocks in New York and the second resource is about fashion and trends on the streets in Berlin and Toronto. For those resource pairs with a high number of co-occurring tags, we do not find that they have high similarity from the topic level.

According to the analysis above, we found that TTR-LDA model can find highly related resource pairs with low common tags, which provides a meaningful method to make resource predictions and tag recommendations.

Second, we would like to discover, for a collection of popular resources, whether or not the activity level of bookmarking can improve the semantic meaning of the resources. That is, when a resource is popular, can its tags provide more accurate semantic information than less-popular and non-popular resources? The results are not as important for users, but are very important for improving text mining. We used perplexity (Blei, D., et al, 2003) to design the experiments. Perplexity is a widely used indicator to show the performance of a statistical model: the lower the perplexity value is, the better a model fits the actual distribution. We found that for popular resources, their tags express more meaningful information than less-popular and non-popular resources. The experiment results can be seen in Table 9:

Table 9: The comparison of Perplexity among popular, less popular and non popular resources

	The number of Topics	Perplexity
1,000 popular resources	300	5723.7438
1,000 less popular resources	300	39253.5928
1,000 non popular resources	300	20896.0291

5 EVALUATION

The macro tag growth of social tagging systems is similar to English corpora and academic articles whose vocabulary growth obeys power-law distributions with an exponent having a sub-linearity along with *tg* (Cattuto et al., 2007). Researchers have found that the range of macro vocabulary growth exponent of traditional English corpora and academic articles is between 0.4 and 0.6 (Harman, 1995). We find the exponent range of social tagging systems to be between 0.8 and 0.9. The micro tag growth of certain resources is similar to the growth of vocabulary in papers and articles, with both having sub-linearity features over time. Based on this we can use similar methods to deal with resources in social tagging systems.

Different social tagging systems also have varying dynamic features. We use Delicious data (with the addition of 2007-2008) to compare our findings with those of Cattuto et al. (2007). We find that the results are consistent with respect to the macro tags growth exponent, exponent of micro tags, taggers growth, average "post length" and resources and tagger activity

probability distribution. The values of tag growth in Flickr and YouTube are not consistent with the values obtained for Delicious.

We also find that the sub-linearity features of popular resources in different tagging systems have a positive relationship with the activity level of taggers. For example, in Delicious, the tagger growth exponents of popular resources converge. Through the average "post length" $\bar{n}$ of posts, we can predicate that the tagger growth exponents of popular resources in Flickr and YouTube converge to a value that is $1/\bar{n}$. We also find that the activity level of taggers has a negative impact on the exponent of macro tag growth, which means that if the taggers are more active, the exponents of macro tag growth may be lower. Understanding the reasons for such a behavior requires further analysis. Our findings confirm that of Suchanek, Vojnovic and Gunawardena (2008) based on a social tagging analysis of 65,000 Delicious bookmarks and a user study of over 4,000 participants, where we all concur that popular resources have more stable tags.

6 CONCLUSION

In this paper, we build up a dynamic model to analyze the features of the three most popular social tagging systems of Delicious, Flickr and YouTube based on large scale tagging data crawled by the UTO crawler. For the social vocabularies, the macro tag growth in the three social tagging systems investigated follow the power-law distribution. When the book-marking activities are accumulated to a certain extent, the growth of new tags shows some regularity (the increasing curve can be fitted by a cubic polynomial), which can be explained as a kind of cognitive process, we used TRR-LDA and perplexity to verify that we can obtain more accurate semantic information from that period.

For tagger activities in Delicious, there is noise at the early stage of tagger growth of ten popular resources, yet after a period of time (when they become popular enough), the curve track of all resources tends to become unified. The tagger activities in all the three applied tagging systems demonstrate normal distribution, while probability distribution of tag growth exponents in Flickr and YouTube shows non-normal distribution. We find that Flickr and Delicious have a similar exponent $\gamma_{U(tg)}$ of tagger growth for popular resources. But YouTube has a bigger post average length (8.2350), which means that the taggers in YouTube provide

more tags per resource, which leads to a lower exponent $\gamma_{U(tg)}$ of tagger growth for certain resources compared with Flickr and Delicious.

Finally, we propose our TTR-LDA model to analyze the tagger-topic-link-tag distribution of the 1,000 most popular resources from 2005 to 2008 on Delicious, and obtain revealing results for the evolutionary features of social tagging topics. We find that a large topic may split into several sub-topics during its evolution. The content of a topic may converge into a relatively stable stage for a period of time, during which the popularity of the topic also tends to be stable, and where a certain group of taggers who have a continuous interest in that topic may be identified.

What we discovered from examining the multi-perspective growth of social tagging vocabulary can be useful for deriving a hybrid or composite indexing schema using the strengths of both folksonomy and traditional indexing. In traditional controlled vocabulary-based indexing, all terms assigned to a document carry more or less equal weight. In social tagging, certain tags become much more popular than others over the entire dataset. This degree of consensus is reached from the reuse/feedback mechanism which enables the folksonomy to be self-regulated. In addition, in a practical sense, understanding how users tag resources help develop various web 2.0 applications for social tagging systems. Moreover, some of the results-for example, growth of number of taggers for various popular resources tend to arrive at similar increasing speed, and growth of number of tags for active taggers shows different normal distribution in different social tagging systems-can be further explored with qualitative research from the perspective of social-technical interactive and cognitive science.

In future work, the TTR-LDA model will be future developed to not only dynamically detect topics from social tagging vocabulary but also to extract clusters or hierarchical structure of topics. This improved model will further reveal the latent semantic structure underlying social tagging vocabulary and open possibilities of connecting controlled vocabulary and social tagging vocabulary, improving tag search, and browsing, building tag recommendation services.

7 ACKNOWLEDGMENTS

Thanks Milojević, Staša for her proof reading and her guidance on this paper.

This work is supported by NIH-funded VIVO project (NIH grant U24RR029822).

Daifeng Li is funded by China National Natural Science Foundation (70971083), the Graduate Innovation Fund of Shanghai University of Finance and Economics (cxjj-2008-330), the 2009 Doctoral Education Fund of Ministry of Education in China (20090078110001) and the NIH VIVO project (uf09179).

Jie Tang is supported by the Natural Science Foundation of China (No. 60703059), Chinese National Key Foundation Research (No. 60933013), and National High-tech R\&D Program (No. 2009AA01Z138).

8 BIBLIOGRAPHY

Ahlgren, P., Jarneving, B., & Rousseau, R. (2003). Requirements for a cocitation similarity measure, with special reference to Pearson's correlation coefficient. *Journal of the American Society for Information Science and Technology*, *54*(6), 550-560.

Altmann, E, G., et al., (2009) Beyond Word Frequency: Bursts, Lulls, and Scaling in the Temporal. *Distributions of Words. PLoS ONE 4(11), e7678.*

Blei, D. M., Ng, A. Y., & Jordan, M. (2003). Latent dirichlet allocation. *Journal of Machine Learning Res., 3*, 993-1022.

Cattuto, C., Baldassarri, A., Servedio, V., & Loreto, V. (2007). Vocabulary growth in collaborative tagging systems. Retrieved May 20, 2010 from http://arxiv.org/abs/0704.3316.

Cattuto, C., Benz, D., Hotho, A., & Stumme, G. (2008). Semantic grounding of tag relatedness in social bookmarking systems. In *Proceedings of the 7th International Semantic Web Conference ISWC2008* (pp. 615–631), Berlin: Springer.

Cattuto, C., et al., (2009) Collective dynamics of social annotation. *PNAS 106(26), 10511-10515.*

Damianos, L., Griffith, J., & Cuomo, D. (2006). Onomi: Social bookmarking on a corporate intranet. Paper presented at the Collaborative Web Tagging Workshop, the *15th International WWW Conference*, Edinburgh, Scotland.

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

Dubinko, M., Kumar, R., Magnani, J., Novak, J., Raghavan, P., & Tomkins, A. (2006). Visualizing tags over time. *ACM Transaction Web, 1*(2), 7.

Golder, S. & Huberman, B. (2006). The structure of collaborative tagging systems. *Journal of Information Science*, 32, 198–208.

Halpin, H., Robu, V., & Shepherd, H. (2007). The complex dynamics of collaborative tagging. In *Proceedings of the 16th International WWW Conference* (pp. 211–220). NY: ACM.

Harman, D. (1995). Overview of the Third Text Retrieval Conference. In *Proceedings of the 3rd Text Retrieval Conference (TREC-3), NIST Special Publication* (pp. 1–19). Darby, PA: DIANE Publishing.

Heymann, P., Ramage, D., Garcia-Molina, H. (2008). Social Tag Prediction. In *Proceedings of the 31stAnnual International ACMSIGIR Conference on Research and Development in Information Retrieval* (pp.531-538). NewYork: ACM Press.

Hotho, A., Jaschke, R., Schmitz, C. & Stumme, G. (2006). Information retrieval in folksonomies: Search and ranking. In Y. Sure and J. Domingue (ed.), *The Semantic Web: Research and Applications* (pp.411-426). New York: Springer.

Kipp, M. E. (2006a). Complementary or Discrete Contexts in Online Indexing: A Comparison of User, Creator and Intermediary Keywords. *Canadian Association for Information Science, Toronto, Ontario, Canada*

Kipp, M. E. I., & Campbell, D. G. (2006b). Patterns and inconsistencies in collaborative tagging systems: An examination of tagging practices. *The American Society for Information Science and Technology, 43*(1), 1-18.

Kipp, M. E. (2006b). Exploring the context of user, creator and intermediate tagging. *IA Summit 2006, Vancouver, BC.*

Kipp, M. E. (2007b). Tagging Practices on Research Oriented Social Bookmarking Sites. *Canadian Association for Information Science, Montreal, Quebec, Canada*

Krestel, R., P. Fankhauser, et al. (2009). Latent dirichlet allocation for tag recommendation. In *Proceedings of the Third ACM Conference on Recommender Systems* (pp. 61-68). New York: ACM Press.

Kumar, R., Novak, J. & Tomkins, A. (2006). Structure and evolution of online social networks. In *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp.611-617). New York: ACM Press.

Li, D., Ding, Y., Qin, Z., Milojević, S., He, B., Yan, E., & Dong, T. (2010). Dynamic features of social tagging vocabulary: Delicious, Flickr, and YouTube. Paper presented at the *2010 International Conference on Advances in Social Networks Analysis and Mining, ASONAM* 2010, 9-11 August 2010, Odense, Denmark.

Li, D, He, B., Ding, Y., Tang, J., Sugimoto, C., Qin, Z., Yan, E., & Li, J (2010) Community-based topic modeling for social tagging. *The 19th ACM International Conference on Information and Knowledge Management (CIKM2010), Oct 26-30, Toronto, Canada.*

Li, X., L. Guo, et al. (2008). Tag-based social interest discovery. In *Proceedings of the 17th International WWW Conference* (pp.675-684). Beijing, China. New York: ACM Press.

Lin, X., Beaudoin, J. E., Bui, Y., & Desai, K. (2006). Exploring characteristics of social classification. Advances in Classification Research, Volume 17; *Proceedings of the17th ASIS&T Classification Research Workshop, Austin, Texas, USA. J.* Furner & J. T. Tennis (Eds.).

Lu, C., Hu, X., Chen, Y., Park, J., He, T., Li, Z. (2010). The topic-perspective model for social tagging systems. *The 16th ACM SIGKDD Conference on Knowledge Discovery and Data Mining. 7.*

Macgregor, G., & McCulloch, E. (2006). Collaborative Tagging as a Knowledge Organisation and Resource Discovery Tool. *Library Review, 55*(5), 291 - 300.

Marlow, C., Naaman, M., boyd, d., & Davis, M. (2006b). Position Paper, Tagging, Taxonomy, Flickr, Article, ToRead. Paper presented at *World Wide Web 2006 (WWW2006): Collaborative Web Tagging Workshop*, Edinburgh, Scotland Retrieved

Michal, R., Griffiths, T., Steyvers, M., & Smyth, P. (2004). The author-topic model for authors and documents. In *Proceedings of the 20th Conference on Uncertainty in Artificial Intelligence* (pp. 487-494). Virginia: Association for Uncertainty in Artificial Intelligence.

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

Morrison P. J. (2008). Tagging and searching: Search retrieval effectiveness of folksonomies on the World Wide Web. *Information Processing & Management, 44*, 1562-1579.

Paolillo, J. (2008). Structure and network in the YouTube core. In Proceedings of the *41st Annual Hawaii International Conference on System Science* (pp. 156–446).Washington, DC: IEEE Computer Society.

Rosen-zvi, M., Griffiths, T., Steyvers, M., & Smyth, P (2004) The author-topic model for authors and documents. *In Proceedings of the 20th Conference on Uncertainty in Artificial Intelligence pp. 487-494. Virginia: AUAI Press.*

Sanguanpong, S., Warangrit, S., & Koht-arsa., K. (2000). *Facts about the thai web*. Retrieved May 20, 2010 from http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.11.745.

Schmitz, C., Hotho, A., Jaschke, R., & Stumme, G. (2006). Mining association rules in folksonomies. *Data Science and Classification*, 4: 261-270.

Serrano, M, A., et al., (2009) Modeling Statistical Properties of Written Text. *PLoS ONE 4(4), e5372.*

Shirky, C. (2005). *Ontology is overrated: Categories, links, and tags.* Retrieved May 20, 2010 from http://shirky.com/writings/ontology overrated.html.

Si, X., & Sun, M. (2009). Tag-LDA for Scalable Real-time Tag Recommendation. *Journal of Information & Computational Science: 6*(1), 23-31.

Suchanek, F. M., Vojnovic, M., & Gunawardena, D. (2008). Social tags: Meaning and suggestions. Paper presented at the *17th ACM Conference on Information and Knowledge Management*, Napa Valley, California, USA.

Smith, T. (2007). Cataloging and You: Measuring the Efficacy of a Folksonomy for Subject Analysis. *18th Workshop of the American Society for Information Science and Technology Special Interest Group in Classification Research,*Milwaukee, Wisconsin, USA. J. Lussky (Ed.),

Tang, J., J. Zhang, et al. (2008). ArnetMiner: extraction and mining of academic social networks. In *Proceeding of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 990-998). Las Vegas, Nevada. USA. New York: ACM.

Tang, J., Jin, R., & Zhang, J. (2008). A Topic Modeling Approach and its Integration into the Random Walk Framework for Academic Search. In *Proceedings of 2008 IEEE International Conference on Data Mining (ICDM'2008)* (pp. 1055-1060). Washington, DC: IEEE Computer Society.

Torunski, L. (2009). *Smart and simple webcrawler*. Retrieved May 20, 2010 from https://crawler.dev.java.net.

Trant, Jennifer (2009) Studying Social Tagging and Folksonomy: A Review and Framework. *Journal of Digital Information 10*(1). Voss, J. (2007).

Veres, C. (2006). The language of folksonomies: What tags reveal about user classification. In C. Kop, G. Fliedl, H. C. Mayr, and E. M´etais (ed.) *NLDB* (pp. 58–69). New York: Springer.

Voss, J. (2007). Tagging, Folksonomy & Co – Renaissance of Manual Indexing. *10th international Symposium for Information Science*.

Weick, K. E., Sutcliffe, K. M., & Obstfeld, D. (2005). Organizing and the Process of Sensemaking. *Organization Science, 16*(4), 409–421.

Xu, S., Bao, S., Fei, B., Su, Z., & Yu, Y. (2008). Exploring folksonomy for personalized search. Paper presented at the Proceedings of the *31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. Singapore.

Zhang, H., Qiu, B., Giles, C. L., Foley, H.C., & Yen, J. (2007). An LDA-based community structure discovery approach for large-scale social networks. In *Proceedings of Intelligence and Security Informatics* (pp. 200-207). Washington: IEEE.