Community-based Topic Modeling for Social Tagging

Daifeng Li[1], Bing He[2], Ying Ding[2], Jie Tang[3], Cassidy Sugimoto[2], Zheng Qin[1], Erjia Yan[2], Juanzi Li[3], Tianxi Dong[1]
[1]School of Information Management and Engineering, Shanghai University of Finance and Economics, Shanghai, China
[2]School of Library and Information Science, Indiana University Bloomington, IN, USA
[3]Department of Computer Science and Technology, Tsinghua University, Beijing, China
ldf3824@yahoo.com.cn, {dingying, sugimoto, eyan, binghe}@indiana.edu, {jietang, ljz}@tsinghua.edu.cn

ABSTRACT
Exploring community is fundamental for uncovering the connections between structure and function of complex networks and for practical applications in many disciplines such as biology and sociology. In this paper, we propose a TTR-LDA-Community model which combines the Latent Dirichlet Allocation model (LDA) and the Girvan-Newman community detection algorithm with an inference mechanism. The model is then applied to data from Delicious, a popular social tagging system, over the time period of 2005-2008. Our results show that 1) users in the same community tend to be interested in similar set of topics in all time periods; and 2) topics may divide into several sub-topics and scatter into different communities over time. We evaluate the effectiveness of our model and show that the TTR-LDA-Community model is meaningful for understanding communities and outperforms TTR-LDA and LDA models in tag prediction.

Categories and Subject Descriptors
H.3.3 [Information Search and Retrieval]: Clustering;
H.2.8 [Database applications]: Data mining

General Terms
Algorithms, Experimentation

Keywords
Topic mining, community detection, social tagging system, TTR-LDA-Community.

1. INTRODUCTION
Social networks have been studied for decades. From a research perspective, these real-world networks display unique properties from the classical random graph model [3] in that most real word networks exhibit three common properties: the small-world property, power-law degree distribution and a high clustering coefficient or transitivity (indicating community structure) [7][8][9]. Intuitively, the heterogeneity of the user groups, the huge quantities of various resources bookmarked, and the variety of interactions among the users provide intrinsic evidence for the existence of user communities. Thus, an important task in network analysis is to detect communities and explore their features, which can improve community-supporting services at the community-level in the context of a social tagging system. Many studies in various disciplines have been devoted to community detection; however, few of them have systematically and quantitatively studied the profiles of those detected communities (for example: the dynamic features of those detected communities; the semantic analysis of those communities which are detected mainly according to link relationship among users).

Recently, statistical topic modeling has been proposed as an unsupervised method to summarize the contents of large document collections. The classic model is called Latent Dirichlet Allocation (LDA) [1]. These models and their extensions use simple surface features such as word occurrences within documents to reveal the semantic content of documents. In this paper, we propose a TTR-LDA-Community model, which is an inferential combination of an extended LDA model and a betweenness-based community detection algorithm. It provides rich, systematic, and quantitative information about the profiles of detected communities.

This paper is organized as follow: Section 2 states related works, Section 3 discusses the method, including the dataset used and the proposed TTR-LDA-Community model. Section 4 presents the results of applying the proposed model to a real world dataset, Section 5 evaluates the methods and results. Section 6 concludes our study.

2. RELATED WORK
Since the introduction of the LDA model [1], various extended LDA models have been used for automatic topic extraction from large-scale corpora. In the context of social tagging systems, where multiple users are annotating resources, the resulting topics reflect a shared view of the document; and the tags of the topics reflect a common vocabulary. As for community detection, the most representative approaches include centrality or betweenness-based approaches and graph partitioning-based approaches. Girvan and Newman extended the betweenness measure to edges and designed a clustering algorithm which gradually removes the edges with the highest betweenness value [4]. This algorithm has been improved through modularity; and the complexity is reduced from $O(m^2n)$ to $O(mdlogn)$ where d is the depth of the dendrogram of the community structure [2]. Many studies provide various models and algorithms for topic mining and community detection; yet, few of them have integrated those models and algorithms, performed topic mining for detected communities, and analyzed how those identified topics change among communities over time. These questions are addressed in this study.

CIKM'10, October 26-30, 2010, Toronto, Ontario, Canada.

3. METHODOLOGY

3.1 Data

The activity of social tagging consists of three major components: tag, tagger and resource. The experimental dataset contains all the triples of these three components and the time and date of their creation on Delicious from 2005 to 2008. In data processing, all taggers were ranked by the number of resources they have bookmarked and the top 50,000 taggers were selected as the sample of taggers. These taggers bookmarked a total of 354,522 web pages, which were sorted by the number of taggers who bookmarked them. The top 10,000 resources were selected as the sample of web pages, associated with which a dominant majority of tagging activities occurred. Thus a co-bookmark network was built in which a connection between two users (within the sample of 50,000 taggers) is created if they bookmarked the same resources (within the sample of 10,000 web pages). In addition, in order to observe the evolution of structure and motif of communities, the time span (2005-2008) was divided into three slices. Table 1 shows the descriptive statistics of the data in the three time slices.

Table 1. Data Statistic

	2005-2006	2007	2008
No. of posts	5,128	22,955	204,129
No. of resources	883	2,094	7,023
No. of tagger	3,965	11,717	34,318
No. of tag	4,701	16,671	67,261

3.2 TTR-LDA-Community model

The TTR-LDA-Community model is an integration of the TTR-LDA model and the Girvan-Newman community detection algorithm, using inference mechanism. The model is illustrated in Figure 1. TTR-LDA is developed based on ACT model [11][12]. It is a three-layer Bayesian model with taggers *tap* in each post *p* as the first layer, tags *t*, and resource *r* as third layer and all the topics denoted as latent variable *z* as the middle layer.

The inference mechanism is used to infer the topic distribution over detected communities. Each community includes a set of taggers, who have a stronger relationship with other taggers within the community than the taggers outside. Based on the taggers' information model, the probability distribution of each tagger over a set of topics is obtained by using the TTR-LDA model while the community structure of taggers is revealed by the community detection algorithm. The two sets of results are further integrated through an inference mechanism. The function of inference mechanism can be described below:

Assuming that we identify a community which contains *p* taggers $\{ tagger_1, tagger_2, tagger_3, ...tagger_p \}$, for $tagger_k$, he/she has a corresponding probability distribution over T topics in $Matrix_{Topic-Tagger}$ as $\{ P_{k1}, P_{k2}, P_{k3}, ...P_{kT} \}$. So for that community, its probability distribution over T topics can be computed as $\{ \frac{\sum_{k=1}^{p} P_{k1}}{p}, \frac{\sum_{k=1}^{p} P_{k2}}{p}, \frac{\sum_{k=1}^{p} P_{k3}}{p}, ... \frac{\sum_{k=1}^{p} P_{kT}}{p} \}$.

3.3 Topic distributions of communities over time

In this section, we not only observe the community and topic distributions of taggers, but also explore how they evolve over time. Taggers and their co-bookmark activities are divided into four time slices (from 2005 to 2008). The fourth time slice (2008), which had the highest number of taggers, is further divided into four periods.

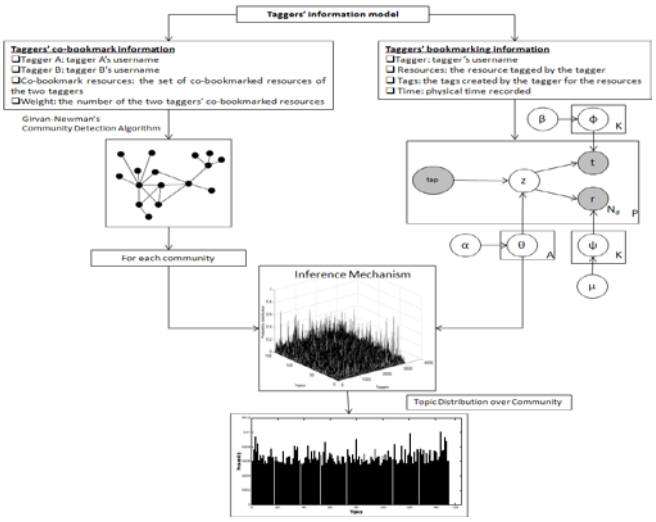

Figure 1. TTR-LDA-Community Model

Results show that the number of users of the top five communities occupies a major proportion in the four years (2005-2008) and the proportion is increasing over time. Posts created during September 2008 to December 2008 are used as training data. A test set of 3,000 is built by sampling 1 out of every 100 posts from these 43,453 distinct taggers and 350,721 different posts. Perplexity is used to identify the number of topics [10], which arrives at the lowest point when the number of topics is 150. The interest model of each tagger in the top five largest communities is then built based on their topic distributions. Taking the largest community in the period as an example, the topic distribution of all the taggers is shown in Figure 2.

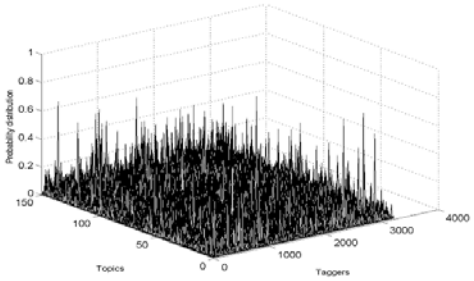

Figure 2. Topic distribution of taggers in the largest community during 2008/10-2008/12.

By using users' interest models and the inference mechanism, a topic distribution of the largest community can be created (Figure3). We can find that the topic distributions in a community are diverse because users' relationships in that community are mainly based on their co-bookmark activities not the similarity of their interest model.

In order to observe the dynamic features of communities, we design an experiment as follows: 1) denote the five largest communities from each time slice in 2008 as *community_i_t* where *t* means the *tth* time slice in 2008 and *i* means the *ith* largest community in *tth* time slice; 2) compute the topic distribution for

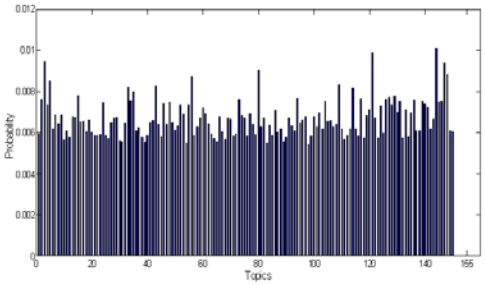

Figure 3. The Topic Distribution in the largest community during 2008/10-2008/12.

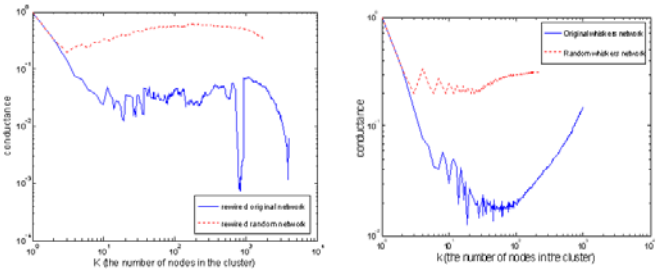

Figure 4. NCP plot

the five communities, which is stored as *model_t_i_Topic(j)*, the probability of *jth* topic in *ith* largest community in the *tth* time slice; 3) obtain the probability distribution of tags that are collected from all the posts generated during the specific time slice; the probability of one tag occurring in a topic shows the level of representativeness of the tag for that topic; 4) sort all the tags according to their probability value in each topic and select the 20 top ranked tags to represent the content of the topics; select the top 5 ranked topics to represent the theme of each community; and 5) analyze the similarity between different communities from different time slice through computing how many tags are shared by the two different communities. More specifically, we compare current time slice with its previous time slice, for example, we compare *community_i_t* with *community_j_t-1(j=1, 2…5)*.

Analysis of the evolutionary line of communities shows that most large-scale communities have a high similarity with *community_1_2* that is mainly on topics related to computer technology. When it comes to the 2nd time slice, the set of topics is divided into two groups: one is related to web technology; the other is about java and business. As for the 4th time slice, the first two communities are purely about web design, and the second two communities are about Web 2.0, social networks and business. The size of communities along evolutionary lines fluctuates over time. For example, the size of the community about social networks in the 3rd time slice (*community_1_3*) is much larger (4,377) than that (521) in the 4th time slice (*community_5_4*).

4. EVALUATION

In this section, the effectiveness of the TTR-LDA-Community model is evaluated, including the quality of detected communities, topic mining, and comparison with other related algorithms.

Community Detection Evaluation

Conductance (from multi-criterion scores) and modularity (from single criterion scores) are used to evaluate the quality of communities detected by the TTR-LDA-Community model [6]. The smaller the value of conductance is, the higher the granularity of a community is. Network community profile (NCP) is used to compute and display the value of conductance for communities [5]. Whiskers networks and rewired networks are adopted as two comparative aspects. Whiskers is defined as the maximal sub graphs that can be detached from the rest of the network by removing a single edge; and a rewired network is a random network that has the same nodes and the same degree distribution as the original network [5]. The conductance of communities of the rewired original network (blue line in the left figure), rewired random network (red dashed line in the left figure), the original whiskers network (blue line in the right figure), and the random whiskers network (red dashed line in the right figure) are calculated and shown in Figure 4.

In Figure 4, compared with the rewired network (left) and the rewired whiskers (right), 1) the original network displays a higher granularity of communities (a lower conductance value); 2) the value of conductance as the function of the size of communities in the original network and the original whiskers present a "V" shape, showing properties of a true large social networks [5]; 3) the original whiskers has the best community granularity (the lowest conductance) between size 10-100; and 4) the best community granularity of rewired original network is around 1000.

Modularity is one of the most widely used methods to evaluate the quality of a division of a network into communities [6]. The modularity value of those detected communities in different time periods is shown in Table 2. The modularity of communities in the four time slices of 2008 is better than that in 2005-2007. This is probably due to the fact that community structure grows mature gradually over time, creating better communities in later years than in earlier years. Meanwhile, modularity of communities in the short-term (four sub periods in 2008) is larger than the long-term (2008). It can be explained that in different time periods, most taggers' bookmarking activities are focused on different domains, so in a certain short-term time period, communities may be quite different from each other. However, when those time periods are merged together, the taggers show different interests in many domains; so the clustering feature within the communities becomes weaker.

Table 2: The modularity values of detected communities in different time periods

Time slice	Modularity	Time slice	Modularity
2005	0.320031	2008, Jan-March	0.797043
2006	0.432471	2008, April-June	0.744134
2007	0.502286	2008, July-Sep.	0.735286
2008	0.524738	2008, Oct. – Dec.	0.645894

Topic Mining Evaluation

Topic distribution of the 1000 most popular resources during 2008-2009 in Delicious is examined. Results show that the most popular topics are about bandslash fiction, fan fiction, and supernatural fiction (the top 3 popular topics). Communities with similar theme are ranked 3rd, 4th, and 5th in size; and the web resources with similar topics are ranked 500-600 of the top 1000 ranked resources in number of taggers associated with them. We also inspect how topics of the top 1,000 resources are distributed in different communities. In Table 3, the *i* in Topic *i(j)* means the *ith* topic in 1000 most popular resource and j denotes the *ith* topic is ranked as *j* in all the 300 topics. The top 20 ranked topics in 1000 most popular resources can be found in 5 largest communities in different time periods. For each community, there

exists at least one topic that is ranked top 10 in 1000 most popular resources (Table 3).

Table 3: The popular topics distribution over communities

	2008 1-3	2008 4-6	2008 7-9	2008 10-12
1st	Topic 153(4) Topic 52(5) Topic 171(11) Topic 14(12)	Topic 153(4) Topic 61(7) Topic 76(16)	Topic 39(61) Topic 236(8) Topic 36(82) Topic 220(9) Topic 61(7)	Topic 236(8) Topic 153(4)
2nd	Topic 100(3)	Topic 153(4) Topic 76(16)	Topic 194(6)	Topic 236(8) Topic 61(7)
3rd	Topic 76(16) Topic 236(8)	Topic 29(1) Topic 100(3)	Topic 194(6) Topic 52(5)	Topic 48(15) Topic 94(6)
4th	Topic 38(52) Topic 48(15) Topic 236(8)	Topic 220(9) Topic 66(102)	Topic 153(4) Topic 52(5) Topic 171 (11) Topic 14(12)	Topic 29(1) Topic 100(3)
5th	Topic 29(1)	Topic 48(15) Topic 194(5)	Topic 220(9) Topic 14(11)	Topic 236(8) Topic 199(13)

Model Evaluation

Topic distributions for each community are obtained respectively from LDA, TTR-LDA model, and TTR-LDA-Community model based on co-bookmark network in a given period (Oct. 2008–Dec. 2008). One resource and five tags are recommended for each post according to the results of three models separately. Taking recommending resources to posts as an example, the web resource that has the highest probability of occurring in a post is recommended to that very post. The results are shown in Table 4.

Table 4. Results of Precision, Recall, F1-measure on dataset from 2008/10 to 2008/12

	Object	Precision	Recall	F1
LDA	Tags for post	0.3502	0.2266	0.2752
TTR-LDA	Tags for post	0.3639	0.2271	0.2797
	Resource for post	0.2690	0.2690	0.2690
TTR-LDA-Community	Tags for post	0.3633	0.2321	0.2809
	Resource for post	0.2873	0.2873	0.2873

In Table 4, only recommendation of tags for each post can be made by using LDA because LDA only provides the probability distribution of posts over topics. The TTR-LDA and TTR-LDA-Community model show significant improvement for recommendation of tags and resources for post in terms of precision, recall and F1-Measure. TTR-LDA and TTR-LDA-Community have slightly improved performance for "tags for post", while TTR-LDA-Community outperforms TTR-LDA on "resource for post".

5. CONCLUSION

This paper proposes the TTR-LDA-Community model, which is an integrated model, which combines TTR-LDA and Community detection using an inference mechanism. By applying this model to Delicious data, the community structure of active taggers, the topic distributions within communities, and the representative taggers, tags, and resources within these communities were observed. Using community detection, the changes in community structure over time were detected. In particular, social tagging communities seem to experience a large intake of newcomers, significantly altering the participant base over time. There also is evidence of a dominance of large communities: the largest of the communities incorporate the majority of participants, although many smaller communities exist. As for topical features of communities, obvious difference exists between communities. Some communities have a core group of topics, while the topic profiles for other communities are varied. Topics may also appear in a few communities simultaneously, and then split into sub-topics and scatter through different communities. In summary, topics seem to be a dynamic feature of communities: emerging, blending, and disappearing over time. The proposed model provides better understanding of the features of communities in social tagging system and provides better opportunities for group recommendation, group prediction and other applications for future research in social tagging. Our future work will focus on building a Dynamic TTR-LDA-Community model, which incorporates evolution of both topics and communities. In addition, future work should explore the dynamic topical features of community profiling.

6. ACKNOWLEDGMENTS

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). Daifeng Li, Zheng Qin is supported by China National Natural Science Foundation (70971083), Graduate Innovation Fund of Shanghai University of Finance and Economics (CXJJ-2008-330), China 211 project fund (211-5-1), 2009 Doctoral Education Fund of Ministry of Education in China (2009007811001).

7. REFERENCES

[1] Blei, D.M., Ng, A.Y., & Jordan, M.I. (2003). Latent Dirichlet allocation. *Journal of Machine Learning Research*: 3 993–1022

[2] Clauset, A., Newman, M. E. J., & Moore, C. (2004). Finding community structure in very large network. *Physical Review E, 70*, 066111

[3] Erdös, P. & Rényi, A.(1959) On random graphs. I, *Publ. Math. Debrecen 6*, 290-*291*.

[4] Girvan, M., & Newman, M. E. J. (2002). Community structure in social and biological networks. *Proceedings of the National Academy of Sciences of the United States of America*, *99*(12), 7821–7826.

[5] Leskovec, J., Lang, K., Dasgupta, A., & Mahoney, M. (2008). Community Structure in Large Networks: Natural Cluster Sizes and the Absence of Large Well-Defined Clusters. arXiv:0810.1355v1.

[6] Leskovec, J., Lang, K., Mahoney, M. (2010). Empirical Comparison of Algorithms for Network Community Detection. In *proceedings of the nineteenth International World Wide Web Conference.* North Caroline, USA.

[7] Milgram, S. (1967). The small world problem, Psychology Today, 2, 60–67

[8] Newman, M. E. J. (2003).The structure and function of complex networks. *SIAM Review* 45, 167–256 .

[9] Newman, M. E. J. (2001). The structure of scientific collaboration networks. *Proceedings of the National Academy of Sciences of the United States of America*, *98*(2), 404-409.

[10] 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.*

[11] Tang, J., Zhang, J., Yao, L., Li, J., Zhang, L., & Su, Z. ArnetMiner: Extraction and Mining of Academic Social Networks. In *Proceedings of the Fourteenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (SIGKDD'2008). pp.990-998

[12] 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.