

An Indirected Recommendation Model for Chinese Microblog

Jianyong Duan, Zheng Dong, Mei Zhang

College of Computer Science, North China University of Technology, Beijing, 100144 China

Email: duanjy@hotmail.com

Abstract: *Microblog is a browser-based platform for web user's information sharing and communication. With the rapidly increasing of microblog population, its recommendation function becomes necessary. This paper proposes the recommendation by the Latent Dirichlet Allocation topic model, which combines the user interests into the model to meet their needs. We also conduct a comparative analysis between indirect and direct recommendation algorithms. The experimental results show that the indirect recommendation is more effective for the micro-blog recommendation.*

Keywords: *Microblog recommendation, the topic model, user interest model.*

1. Introduction

Microblog is a popular social media [1]. It has been accepted by the majority of web users because of its advantages to convenience, low cost and high efficiency. With Sina micorblog, for example, until December 2013, its number of monthly active users reached 129.1 million and the number of daily active users reached 61.4 million in China. At the same time, it has also gradually accumulated abundant information. How to effectively choose the needed information always confuses the web users [2]. Thus, microblog recommendation is urgent for web user [3].

In this paper, we introduce the Latent Dirichlet Allocation (LDA) for microblog topic model construction [4]. The information of microblog is scattered into topics by this model. Then the recommendation system effectively accumulates the weights of user interests and found the users' interests.

2. Related work

There is already some research in the area of microblog recommendation [5, 6], such as user-related recommendation and tag-based recommendation. User-related recommendation allows users to read more microblogs from their friends. Tag-based recommendation lets the system understand and serve users.

The difficulties of recommendation are also explored. Firstly, most microblogs have no clear topics [7, 8]. Those microblogs often describe the user's own mood or interactive content. Secondly, user interest is always changing [9]. The Microblog is a platform for rapid information dissemination. Users easily switch their interests by their browsed information. Thus, user's behavior is difficult to capture [10]. Due to limited content of microblog post, user may stay only a few seconds in one topic, it is difficult to capture the user preference for certain topics [11]. Moreover, most users rarely comment on the topics. The system cannot effectively capture user's interest.

LDA topic model as an effective probabilistic semantic analysis model, as it is widely used into topic detection, clustering, recommendation and other issues[12,13]. In the LDA topic model, there are three layers including documents, topics and words. The document consists of several topics. The document can be expressed as the probability distribution of topics [14]. Every topic consists of a number of words. It can be expressed as the distribution of a number of words.

3. User topic model construction

3.1. The LDA topic model

The LDA topic model is a kind of hierarchical Bayesian model [15]. It is composed of three levels, such as documents, topics and words. A document consists of multiple topics with probabilities. A topic consists of multiple words with probabilities. Then the distribution of words in the document is represented as

$$(1) \quad p(\text{word} | \text{document}) = \sum_{\text{topic}} p(\text{word} | \text{topic}) \times p(\text{topic} | \text{document}).$$

Assuming that there are m documents and n independent words in the document set D . Then each topic (also as theme) can be expressed as an n -dimensional vector φ , which is subject to the Dirichlet distribution β . If there are x topics, each document can be expressed as an x -dimensional vector θ , which is subject to parameters α of the Dirichlet distribution.

LDA topic model generation process is shown in Fig. 1. When give one document, the system selects one topic from the document. Then the system selects one word from the topic. This process repeats. Finally a document LDA is generated.

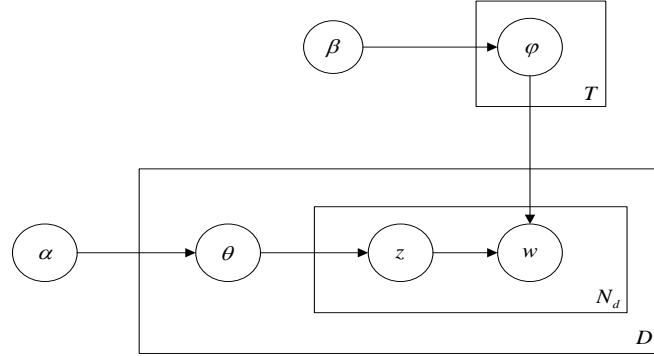


Fig. 1. LDA topic model

3.2. User personalized interest combination

The LDA model is composed of three levels including document layer, topic layer and word layer. For the convenience of describing the interest of users, their personalized interests are added into this model [16, 17], it is the key issue of users recommendation.

In our LDA topic model, words layer as $W = \{w_1, w_2, w_3, \dots, w_n\}$, which is the set after removing stop words; topic layer as $T = \{z_1, z_2, z_3, \dots, z_t\}$, each topic is a set of words of the multinomial distribution, which is subject to $\varphi_i = (q_{i,1}, q_{i,2}, q_{i,3}, \dots, q_{i,n})$, $\sum_{j=1}^n q_{i,j} = 1$, and $q_{i,j}$ represents the probability of a word w_j in the topic z_i ; document layer as $D = \{\theta_1, \theta_2, \theta_3, \dots, \theta_m\}$, each document is a set of topics of the multinomial distribution, which is subject to $\theta_d = (p_{d,1}, p_{d,2}, p_{d,3}, \dots, p_{d,n})$, $\sum_{j=1}^n p_{d,j} = 1$, where $p_{i,j}$ represents the probability of a topic z_j in the document d .

The user interest layer is added into the LDA model as the set $U = \{u_1, u_2, u_3, \dots, u_y\}$. Each user is based on a cumulative variable θ , expressed as

$$(2) \quad u_i = \left(\sum_{d=S} p_{d,1}, \sum_{d=S} p_{d,2}, \sum_{d=S} p_{d,3}, \dots, \sum_{d=S} p_{d,n} \right),$$

where S is the number of documents which are visited by users.

3.3. Clustering interest topics

For avoiding repeated recommendation, we cluster the similar interest topics and group them as single topic [18]. It improves the recommendation diversity. K -Means++ algorithm is used to cluster [19]. It is unsupervised machine learning, and also has a better performance than K -Means algorithm.

Assuming that the topic set is $T = \{z_1, z_2, z_3, \dots, z_m\}$, and k initial centroid of the optimized set is $P = \{p_1, p_2, p_3, \dots, p_k\}$. Then our clustering steps as following:

Step 1. Find the nearest centroid p_i from each topic as

$$(3) \quad \text{tmp}_i = \min_j \|z_i - p_j\|^2, \quad i \in \{1, 2, 3, \dots, m\}, \quad j \in \{1, 2, 3, \dots, k\}.$$

Step 2. According to the new cluster results, the system relocates the centroid p_j as

$$(4) \quad p_j = \frac{\sum_{i=1}^m 1\{\text{tmp}_i = j\} z_i}{\sum_{i=1}^m 1\{\text{tmp}_i = j\}}.$$

Step 3. Repeat the process described above until it has no change. The centroid $\{\text{tmp}_1, \text{tmp}_2, \text{tmp}_3, \dots, \text{tmp}_m\}$ collection is the final result, where $\text{tmp}_i \in \{1, 2, 3, \dots, k\}$ represents the number of centroids.

4. Recommendation process

The recommended procedure consists of three steps. (I) choose the time interval. (II) classify the user interest. (III) calculate the weights of recommendation contents.

4.1. Choose the time interval

The user interest changes dynamically over time and is subject to certain distribution. We select two time intervals, such as natural time interval and microblog operating time interval [20]. Natural time interval means any time period of our daily life. Microblog operating time interval means the time period of users' forwarding, posting and other behaviors.

Assuming that every interval has equal period time, in the interval t_1 , the user's interest is as

$$(5) \quad u_{t_1} = \left(\sum_{d=S_{t_1}} P_{d,1}, \sum_{d=S_{t_1}} P_{d,2}, \sum_{d=S_{t_1}} P_{d,3}, \dots, \sum_{d=S_{t_1}} P_{d,n} \right),$$

and in the continued interval t_2 , the user's interest is as

$$(6) \quad u_{t_2} = \left(\sum_{d=S_{t_2}} P_{d,1}, \sum_{d=S_{t_2}} P_{d,2}, \sum_{d=S_{t_2}} P_{d,3}, \dots, \sum_{d=S_{t_2}} P_{d,n} \right),$$

where $\sum_{i=S_t} P_{i,j}$ means the cumulative distribution of related topic j during the time interval t . The increment of the user interest is expressed as

$$(7) \quad \begin{aligned} u_{t_2} - u_{t_1} &= (\Delta \sum_{d=\Delta S} P_{d,1}, \Delta \sum_{d=\Delta S} P_{d,2}, \Delta \sum_{d=\Delta S} P_{d,3}, \dots, \Delta \sum_{d=\Delta S} P_{d,n}) = \\ &= (\Delta p_1, \Delta p_2, \Delta p_3, \dots, \Delta p_n), \end{aligned}$$

where

$$(8) \quad \Delta \sum_{d=\Delta S} P_{d,j} = \sum_{d=S_{t_2}} P_{d,j} - \sum_{d=S_{t_1}} P_{d,j}.$$

4.2. Classify the user interest

Assuming that the set of user interest is $C = \{c_1, c_2, c_3, \dots, c_y\}$, where C represents a set which is composed by a number of topics c_i .

Suppose the weight set is $G = \{g_1, g_2, g_3, \dots, g_y\}$, where g_y represents the corresponding weight of topic c_i . The user interest tendency is

$$(9) \quad g_i = \frac{\sum_{j=1}^n f(i, j)}{\sum_{j=1}^n h(i, j)},$$

where

$$(10) \quad f(i, j) = \begin{cases} \Delta p_j, & z_j \in c_i, \\ 0, & z_j \notin c_i, \end{cases} \quad h(i, j) = \begin{cases} 1, & z_j \in c_i, \\ 0, & z_j \notin c_i, \end{cases}$$

$$(11) \quad g_{\max} = \max\{g_1, g_2, g_3, \dots, g_y\}.$$

The user interest tendency is classified into c_{\max} which corresponds to the g_{\max} .

4.3. User recommendation

According to the user interest tendency, we propose two methods including direct and indirect recommendation method respectively.

4.3.1. Direct recommend

This method directly recommends microblogs to users. Both common interests in the same category and personalized interest of user in the interval t are combined into this model.

Firstly, user interest tendency is converted to the multinomial distribution as

$$(12) \quad p_{\max} = \{p_1, p_2, p_3, \dots, p_n\} \quad \begin{cases} p_j = \Delta p_j, & z_j \in c_{\max}, \\ p_j = 0, & z_j \notin c_{\max}, \end{cases}$$

and it is transformed into binomial distribution as

$$(13) \quad p_u = \left\{ \frac{p_1}{\sum_{i=1}^n p_i}, \frac{p_2}{\sum_{i=1}^n p_i}, \frac{p_3}{\sum_{i=1}^n p_i}, \dots, \frac{p_n}{\sum_{i=1}^n p_i} \right\}.$$

Then the system calculates the distribution of the distance between microblog topic distribution $\theta_d = (p_{d,1}, p_{d,2}, p_{d,3}, \dots, p_{d,n})$ in the category and user interest distribution p_u .

We use the Kullback-Leibler (KL) divergence to compute the probability distribution [21], as (14), the same way for $D_{\text{KL}}(p_u \parallel \theta_d)$ and $D_{\text{KL}}(\theta_d \parallel p_u)$:

$$(14) \quad D_{\text{KL}}(P \parallel Q) = \sum_i P(i) \ln \frac{P(i)}{Q(i)}.$$

The distribution similarity of p_u and θ_d is calculated as

$$(15) \quad L = \frac{2}{D_{\text{KL}}(\theta_d \parallel p_u) + D_{\text{KL}}(p_u \parallel \theta_d)}.$$

Then system ranks those interested microblogs in descending order as a recommendation.

4.3.2. Indirect recommendation

The keywords in topics are used to recommend instead of microblog in this method. Then the system recommends related microblogs for users by their selected keywords. During the recommendation process, users adjust their intents gradually.

Firstly the system calculates weights of topics. The interest set is defined as $C = \{c_1, c_2, c_3, \dots, c_y\}$. For each topic of microblog, it generates probability weights, and let c_i be the weight of topic as

$$(16) \quad c_i = \left(\sum_{d=S_{\text{all}}} p_{d,1}, \sum_{d=S_{\text{all}}} p_{d,2}, \sum_{d=S_{\text{all}}} p_{d,3}, \dots, \sum_{d=S_{\text{all}}} p_{d,n} \right),$$

where n is the topic number of the classification containing.

Secondly, the frequency of keyword in one topic is described as

$$(17) \quad z_t = (w_{1,\text{count}}, w_{2,\text{count}}, w_{3,\text{count}}, \dots, w_{m,\text{count}}).$$

Assuming that one topic is z , it is $\sum_{d=S} p_{d,i} > 0$ in the u set. And it is included

in c_i set as $\sum_{d=S_{\text{all}}} p_{d,j}$. Its weight is calculated as

$$(18) \quad \text{weight} = \text{MaxCount}(z) \times \sum_{d=S_{\text{all}}} p_{d,j},$$

where

$$(19) \quad \text{MaxCount}(z) = \max\{w_{1,\text{count}}, w_{2,\text{count}}, w_{3,\text{count}}, \dots, w_{m,\text{count}}\}.$$

After the above calculation for each subject of u set, the system gets $\text{Weight} = \{\text{weight}_1, \text{weight}_2, \text{weight}_3, \dots, \text{weight}_n\}$. Each item in the collection represents weight values of words. Then the system ranks several keywords by descending order as a recommendation results.

When users click on a recommendation word, the system finds the appropriate topic category. Assuming that $Z_{\text{result}} = \{z_1, z_2, z_3, \dots, z_m\}$ is the word set of topic. Then the system finds related microblogs which contains Z_{result} .

The topic of microblog is described as $\theta_d = (p_{d,1}, p_{d,2}, p_{d,3}, \dots, p_{d,n})$. For any microblog θ_d , if one topic is $z_i \in Z_{\text{result}}$ and $p_{d,j} > 0$ for θ_d in the z_i , the z_i topic will be accumulated as a result of recommended weights $\text{weight}_{\text{result},d}$. Finally, the system ranks these microblogs as recommendation results.

5. Experimental results

5.1. Data set

We grabbed 10,051 user relationships and about 600,000 microblog documents during May 2014 to November 2014. Their topics include sports, science, current events, people's lives, entertainment and other categories. Their data in these documents has been cleaned as following steps below.

Step 1. Delete the posts which contain less than 80 characters.

Step 2. Classify the user types. According to the post number of microblog, we can category them into personal or public microblogs. For instance, more than 10 posts a day is usually considered as public users.

Step 3. Remove those inactive users who post, forward or comment microblog posts less than two times a week.

After data cleaning, we collect 131 original users which associate with 4,876 microblog posts, and 215 public users which associate with 66,901 microblog posts.

5.2. Measures

5.2.1. Perplexity

The perplexity is used to adjust the parameters of the LDA topic model [22]. The smaller the perplexity is, the stronger its generalization ability is:

$$(20) \quad \text{perplexity}(D_{\text{text}}) = \exp \left| \frac{\sum_{i=1}^M \log p(w_i)}{\sum_{i=1}^M N_i} \right|,$$

where D_{text} represents the test document collection, M represents the number of microblog posts in the test set, $p(w_i)$ is generation probability of a word in one document, N_i is the total number of words in the collection.

LDA topic model needs the topic number $T \left(\alpha = \frac{50}{T} \right)$ because different T values lead to different perplexities. In the experiment, topic numbers are ranged from 100 up to 1000, their trends as it is shown in Fig. 2.

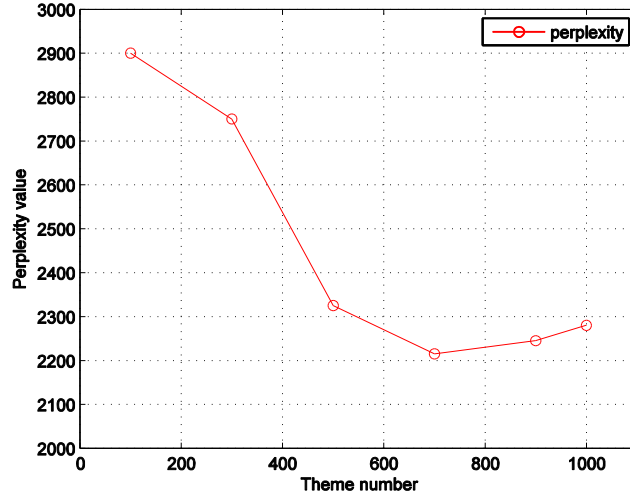


Fig. 2. The changes of perplexity

According to Fig. 2, when the topic number is 700, the perplexity value is relatively stable. Thus 700 is our value and they are divided into 50 categories by K-Means++.

5.2.2. Test measure

We use the recall and precision rate to evaluate the experimental results. Recall rate is the ratio of N_r , which is the number of actual recommendations, and N_R , which is the number of actual meeting the user needs as

$$(21) \quad \text{recall} = \frac{N_r}{N_R}.$$

Precision rate is the ratio of N_r and N_U which is the number of system recommended as

$$(22) \quad \text{precision} = \frac{N_r}{N_U}.$$

5.3. Results and analysis

We select the different time intervals for experiments. The first group of experiment is designed for natural time interval, such as (1) One day as a time T_1 , (2) Two days as a time T_2 ; the second group of experiment is designed for microblog operating time interval, as (3) Forwarding two microblog posts as a time T_3 , (4) Forwarding four microblog posts as a time T_4 .

In the experiment, 130 ordinary users are divided into the 10 test groups and 13 users per group. 10 groups of users are tested by direct recommendation method, their results are as it is shown in Table 1 for T_1 and T_2 , and Table 2 for T_3 and T_4 .

Table 1. Direct recommendation results about T_1 and T_2

User group	$\Delta t=T_1$		$\Delta t=T_2$	
	Average recall rate	Average precision rate	Average recall rate	Average precision rate
Group 1	0.2296	0.2379	0.2081	0.1290
Group 2	0.2931	0.2189	0.2600	0.1930
Group 3	0.2882	0.2121	0.2111	0.1250
Group 4	0.2276	0.2176	0.2857	0.2077
Group 5	0.2428	0.2456	0.2053	0.1333
Group 6	0.2650	0.1931	0.2889	0.1739
Group 7	0.2290	0.1081	0.2930	0.1600
Group 8	0.1952	0.1729	0.1900	0.1538
Group 9	0.2352	0.1176	0.2976	0.1481
Group 10	0.2500	0.2082	0.2100	0.1429

From Table 1, we know that the results are similar for two kinds of time intervals. It means that the time interval is not deciding factor for recommendation, especial for their behaviours including comment and forwarding.

Table 2. Direct recommendation results about T_3 and T_4

User group	$\Delta t=T_3$		$\Delta t=T_4$	
	Average recall rate	Average precision rate	Average recall rate	Average precision rate
Group 1	0.2857	0.2077	0.2714	0.1818
Group 2	0.3250	0.2111	0.2667	0.1875
Group 3	0.3143	0.2212	0.3179	0.2110
Group 4	0.2053	0.2333	0.2622	0.1935
Group 5	0.2181	0.2290	0.3556	0.1925
Group 6	0.3151	0.2613	0.2526	0.2167
Group 7	0.2716	0.1867	0.2389	0.1963
Group 8	0.2982	0.2724	0.2571	0.2846
Group 9	0.3429	0.1915	0.2152	0.1520
Group 10	0.2421	0.2113	0.2724	0.1582

From Table 2 known, the forwarding behaviour has more influence for recommendation than time interval. It indicates that user interests are more easily observed by their behaviours.

The Table 1 and Table 2 are results of direct recommendation. Their precisions are relatively low because the users may be interest in many topics during the time interval, while the system only selects one major interest topic to recommend.

Thus, we conduct the indirect recommendation experiment. Firstly the system provides users with several keywords which relate their interests. Then it recommends to users related topics, based on their interest keywords. In the test, we select 10 keywords for per user during certain interval. And compute the recall and precision as in Table 3 and Table 4.

Table 3. Indirect recommendation results about T_1 and T_2

User group	$\Delta t=T_1$		$\Delta t=T_2$	
	Average recall rate	Average precision rate	Average recall rate	Average precision rate
Group 1	0.3296	0.2521	0.2985	0.2823
Group 2	0.3101	0.2985	0.3251	0.3001
Group 3	0.3258	0.2781	0.2191	0.2858
Group 4	0.3588	0.3001	0.3024	0.2425
Group 5	0.3589	0.2847	0.3208	0.2921
Group 6	0.3025	0.2813	0.3114	0.2471
Group 7	0.3015	0.2961	0.3058	0.2514
Group 8	0.3521	0.2512	0.2814	0.2517
Group 9	0.3111	0.2528	0.3005	0.2617
Group 10	0.3481	0.2747	0.2828	0.2759

Table 4. Indirect recommendation results about T_3 and T_4

User group	$\Delta t=T_3$		$\Delta t=T_4$	
	Average recall rate	Average precision rate	Average recall rate	Average precision rate
Group 1	0.3578	0.3077	0.3845	0.2813
Group 2	0.36814	0.2821	0.3422	0.2721
Group 3	0.3633	0.2912	0.3331	0.2878
Group 4	0.3878	0.3123	0.3281	0.2673
Group 5	0.4001	0.2901	0.3551	0.2581
Group 6	0.4147	0.3213	0.3811	0.2877
Group 7	0.3854	0.3100	0.3111	0.2532
Group 8	0.4111	0.3133	0.3418	0.2561
Group 9	0.3858	0.3045	0.3229	0.2566
Group 10	0.3958	0.3021	0.3671	0.2913

From the experimental results, we know that the indirect recommendation is better than direct recommendation. The reason is that one microblog document may contain some topics, and the interested topics of users cannot be in full accord with the microblogs.

In the indirect recommendation method, the topic is represented by a group of keywords. It is a kind of flexible method with larger coverage for topic. Then the system requires users selecting keywords to represent their interests. The user selected keywords indicate their intentions. The system easily meets the true needs of users and improves the performance of recommendation.

6. Conclusion

This paper proposes the microblog recommendation by the LDA topic model. We compare two recommendation methods, such as direct and indirect recommendation methods. Experimental results show that indirect recommendation is better than direct recommendation.

In the future, we will incorporate the user's behaviour into recommendation model. In the topic classifications, we will analyze the correlation among the topics in order to allow users to have a better experience.

Acknowledgements: This work is supported by the National Science Foundation of China (Grant No 61103112), Social Science Foundation of Beijing (Grant No13SHC031) and Beijing Young talent plan (Grant No CIT\&TCD201404005).

References

1. Sun, L., Y. Liu, Q.-A. Zeng, F. Xiong. A Novel Rumor Diffusion Model Considering the Effect of Truth in Online Social Media. – *International Journal of Modern Physics*, Vol. **26**, 2015, No 7, pp. 1-20.
2. He, Y., J. Tan. Study on Sina Micro-Blog Personalized Recommendation Based on Semantic Network. – *Expert Systems with Applications*, Vol. **42**, 2015, pp. 4797-4804.
3. Zhou, X., S. Wu, C. Chen, G. Chen. Real-Time Recommendation for Microblogs. – *Information Sciences*, Vol. **279**, 2014, pp. 301-325.
4. Blei, D. M., A. Y. Ng, M. I. Jordan. Latent Dirichlet Allocation. – *Journal of Machine Learning Research*, Vol. **3**, 2003, No 4, pp. 993-1022.
5. Liu, Q., H. Ma, E. Chen, H. Xiong. A Survey of Context-Aware Mobile Recommendations. – *International Journal of Information Technology and Decision Making*, Vol. **12**, 2013, No 1, pp. 139-172.
6. Pan, Y., L. Luo, D. Liu. How to Recommend by Online Lifestyle Tagging. – *International Journal of Information Technology and Decision Making*, Vol. **13**, 2014, No 6, pp. 1183-1209.
7. Zhao, B., Z. Zhang, W. Qian, A. Zhou. Identification of Collective Viewpoints on Microblogs. – *Data and Knowledge Engineering*, Vol. **87**, 2013, pp. 374-393.
8. Gao, K., H. Xu, J. Wang. A Rule-Based Approach to Emotion Cause Detection for Chinese Microblogs. – *Expert Systems with Applications*, Vol. **42**, 2015, pp. 4517-4528.
9. Allen, S. M., M. J. Chorley, G. B. Colombo, E. Jaho, M. Karaliopoulos, I. Stavrakakis, R. M. Whitaker. Exploiting User Interest Similarity and Social Links for Micro-Blog Forwarding in Mobile Opportunistic Networks. – *Pervasive and Mobile Computing*, Vol. **11**, 2014, pp. 106-131.
10. Li, H., J. Yan, H. Weihong, D. Zhaoyun. Mining User Interest in Microblogs with a User-Topic Model. – *China Communications*, Vol. **11**, 2014, No 8, pp. 131-144.
11. Bosch, H., D. Thom, F. Heimer. Scatterblogs: Real-Time Monitoring of Microblog Messages through User-Guided Filtering. – *IEEE Transactions on Visualization and Computer Graphics*, Vol. **19**, 2013, No 12, pp. 2022-2031.
12. Jeong, D.-H., M. Song. Time Gap Analysis by the Topic Model-Based Temporal Technique. – *Journal of Informetrics*, Vol. **8**, 2014, pp. 776-790.
13. Haidar, M. A., D. O'Shaughnessy. Unsupervised Language Model Adaptation Using Lda-Based Mixture Models and Latent Semantic Marginals. – *Computer Speech and Language*, Vol. **29**, 2015, No 1, pp. 20-31.
14. Vulic, W., D. Smet, J. Tang, M.-F. Moens. Probabilistic Topic Modeling in Multilingual Settings: An Overview of its Methodology and Applications. – *Information Processing and Management*, Vol. **51**, 2015, pp. 111-147.
15. Li, X., J. Ouyang, X. Zhou. Supervised Topic Models for Multi-Label Classification. – *Neurocomputing*, Vol. **149**, 2015, pp. 811-819.
16. Jie, L., L. Yun, Z. Zhen-Jiang, G. C. Ni. Personalized Recommendation via an Improved NBI Algorithm and User Influence Model in a Microblog Network. – *Physica A: Statistical Mechanics and its Applications*, Vol. **392**, 2013, No 19, pp. 4594-4605.

17. Tang, J., Z. Liu, M. Sun, J. Liu. Portraying User Life Status from Microblogging Posts. – Tsinghua Science and Technology, Vol. **18**, 2013, No 2, pp. 182-195.
18. Ravikumar, S., R. Balakrishnan, S. Kambhampati. Ranking Tweets Considering Trust and Relevance. – In Information Integration in the Web'2012, pp. 1-4.
19. Arthur, D., S. Vassilvitskii. K-Means++: The Advantages of Careful Seeding. – Tech. Rep., Stanford University, 2007.
20. Chen, X., R. Geng, S. Cai. Predicting Microblog Users' Lifetime Activities a User-Based Analysis. – Electronic Commerce Research and Applications, 2014.
<http://dx.doi.org/10.1016/j.elerap.2014.06.001>
21. Kullback, S., R. Leibler. On Information and Sufficiency. – Annals of Mathematical Statistics, Vol. **22**, 1951, No 1, pp. 79-86.
22. Chang, J., J. Boyd-Graber, C. Wang, S. Gerrish, D. M. Blei. Reading Tea Leaves: How Humans Interpret Topic Models. – NIPS, 2009, pp. 1-9.