ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering
*(An ISO 3297: 2007 Certified Organization)*
Website: www.ijircce.com
Vol. 5, Issue 5, May2017

# Temporal Topic Model for Friend Recommendation System

Jayashri Vijay Thorat

Master of Engineering Student, Department of Computer Engineering, ICOER, Wagholi, Pune, India

**ABSTRACT**: Microblogging e.g. Twitter as a brand new variety of on-line communication within which users observe their daily lives, publish opinions or share info by short posts, has become one in all the foremost fashionable social networking services nowadays, that makes it doubtless an oversized info base attracting increasing attention of researchers within the field of information discovery and data processing. During this paper, we tend to conduct a survey regarding existing analysis on info extraction from microblogging services and their applications, and so address some promising future works. We tend to specifically analyze three styles of information: personal, social and travel data.Microblogging users reflect their hobbies or interest and the essential keywords in the messages show their main focus to a huge extent, we can find users' preferences by investigating the user generated contents. Besides, user's hobbies, interest are not static; despite what might be expected, they change as time passes by.

**KEYWORDS**: K-means Algorithm, FLDA,Microblogging, Twitter,Keyword Extraction, Keyword Clustering, Computer Mediated Chat (CMC).

## I. INTRODUCTION

In the current era, individuals are getting additional communicative through enlargement of services and multi-platform applications, i.e., the thus known as internet two.0 that establishes social and cooperative backgrounds. They unremarkably use varied means that together with Blogs to share the diaries, RSS feeds to follow the most recent info of their interest and laptop mediate Chat (CMC) applications to carry duplex communications. Microblogging is one among the foremost recent merchandise of CMC, within which users refer their daily lives, publish opinions or share info by short posts. it had been 1st called Tumblelogs on April twelve, 2005, then came into larger use by the year 2006 and 2007, once such services as Tumblr and Twitter arose. In keeping with social statistics, there have been 111 microblogging sites internationally in could 2007. Among the foremost notable microblogging services these days square measure Twitter, Tumblr, Plurk and Chinese SinaWeibo, to call a couple of. As a well-developed and widely-used microblogging service, Twitter has spawned nice analysis interest recently. Therefore, during this paper, we tend to specifically specialize in Twitter to review the task of data extraction from microblogging services.

Twitter provides its users a strict limit of a hundred and forty characters per posting (often known as tweet) for broadcasting something they need. Twitter users will take different users' tweets by following specific users a bit like most on-line social networking services do, like Facebook and MySpace. However, this follower-and-follower relationship in Twitter needs no reciprocation. That is, the user being followed needn't follow back. On receiving a tweet, users will treat that tweet or re-tweet (identified by `RT') after they it of some interest, that empowers a tweet to be visible outside its original one-degree subscribing network. to boost its race feature, Twitter additionally predestines a special mark-up vocabulary: `@' followed by a username symbol to deal with that exact user or to initiate a directed oral communication, and `#' followed by a sequence of characters to represent hash-tags, that add further context to tweets and facilitate straightforward search of tweets that contain similar hash-tags. Since its birth in Oct 2006, Twitter has become one amongst the foremost notable social networking and microblogging services nowadays, \with over three hundred million users as of 2011, generating over three hundred million tweets and handling over one.6 billion search queries per day. The apace growing worldwide quality makes twitter probably an outsized data base attracting increasing attention of researchers within the field of information discovery and data processing. Actually, data detection from Twitter has long been a hot analysis topic within the internet Community recently.

However, it's value mentioning that extracting helpful data from Twitter may be a advanced task, quite merely applying the standard data extraction technologies that are tried thriving within the internet corpus or different social networking sites to the Twitter context. Twitter has some distinct characteristics that build the data extraction method more difficult. As an example, in contrast to internet documents or blogs, the postings on Twitter area unit continuously short because of the 140-character length limit, thus users will not take an excessive amount of thinking before creating a post. This usually leads tweets to be creaking, ill-formed, and choked with abbreviations, symbols and misspellings. As a consequence, ancient human language technology tools like POS taggers or Named Entity Recognizers (NERs) cannot be applied on to Twitter. Withal, these options additionally bring several new opportunities to researchers on Twitter. As an example, the length-limitation of tweets makes it easier to broadcast a posting, so successively creating the data contained on the Twitter platform underclassman and additional real times. This opens probabilities of victimisation tweets to predict returning trends or notice current events. Besides, in contrast to different social networking sites like Facebook and MySpace, the subsequent network on Twitter is directional instead of reciprocal. In different words, users' requests to take others don't need the target users' approval. Then during this state of affairs, a way to guarantee privacy has additionally become a promising whereas difficult analysis topic.

The purpose of our work is to conduct a survey regarding existing analysis on data extraction from Twitter, likewise as its applications in real world, and advocate some promising future works to researchers fascinated by this field. we have a tendency to specifically analyse 3 varieties of information: Personal data that's typically contained during a user's profile, as well as demographic options like age, gender, quality and residential address; and alternative options as well as health standing, orientation, business information, user interest then on.

## II. LITERATURE SURVAY

**1) Characterizing Microblogs with Topic Models**
**Authors: Daniel Ramage, Susan Dumais, Dan Liebling**
**Description:**
As microblogging grows in quality, services like Twitter are coming back to support military operation desires on top of and on the far side their ancient roles as social networks. However most users' interaction with Twitter remains primarily targeted on their social graphs, forcing the customarily inappropriate conflation of "people I follow" with "stuff i would like to scan." we tend to characterize some info desires that the present Twitter interface fails to support, and argue for higher representations of content for resolution these challenges. We tend to gift a ascendible implementation of a partly supervised learning model (Labeled LDA) that maps the content of the Twitter feed into dimensions. These dimensions correspond roughly to substance, style, status, and social characteristics of posts. We tend to characterize users and tweets mistreatment this model, and gift results on 2 info consumption oriented tasks.

**2) Paper Name: A recommender system based on tag and time information for social tagging systems**
**Authors: Nan Zheng, Qiudan Li**
**Description:**
Recently, social tagging has become progressively current on the net that provides a good approach for users to arrange, manage, share and rummage around for numerous types of resources. These tagging systems supply immeasurable helpful data, like tag, Associate in nursing expression of user's preference towards an explicit resource; time, a denotation of user's interest's drift. As data explosion, it's necessary to advocate resources that a user would possibly like. Since cooperative filtering (CF) is aimed to produce customized services, the way to integrate tag and time data in CF to produce higher customized recommendations for social tagging systems becomes a difficult task. During this paper, we have a tendency to investigate the importance and quality of tag and time data once predicting users' preference and examine the way to exploit such data to create a good resource-recommendation model. we have a tendency to style a recommender system to comprehend our machine approach. Also, we have a tendency to show by trial and error exploitation knowledge from a real-world dataset that tag and time data will well specific users' style and that we conjointly show that higher performances are often achieved if such data is integrated into CF.

**3) Paper Name: Comparing Twitter and Traditional Media using Topic Models**
**Authors: Wayne Xin Zhao1, Jing Jiang2, Jianshu Weng2, Jing He1, Ee-Peng Lim2, Hongfei Yan1 and Xiaoming Li**
**Description:**Twitter as a brand new type of social media will probably contain much helpful data; however content analysis on Twitter has not been well studied. Specifically, it's not clear whether or not as Associate in nursing data

source Twitter will be merely thought to be a quicker news feed that covers mostly a similar data as ancient print media. During this paper we through empirical observation compare the content of Twitter with a standard news medium, any Times, victimisation unsupervised topic modelling. We use a Twitter-LDA model to find topics from a sample distribution ofthe entire Twitter. We tend to then use text mining techniques to check theseTwitter topics with topics from any Times, taking into thought topic classes and kinds. We tend to additionally study the relation between the proportions of narrow tweets and re-tweets and topic classes and types. Our comparisons show attention-grabbing and helpful findings for downstream IR or DM applications.

## III.EXISTING SYSTEM

In existing ancient recommendation systems, several prediction algorithms, like the singular worth decomposition (SVD) based mostly formula, area unit then conducted directly on these distributed matrices to fill out the missing parts. Considering the cold-start downside, before prediction and recommendation, we tend to optimize the tags of microblog victimization the interest evolution model and initialize user preference with the colds tart downside by social tag prediction.

Most of the current friend suggestions mechanism depends on pre-existing user relationships to select friend candidates like friend of friend i.e. mutual friends.

**Disadvantages of Existing System:**

1) The SVD based mostly formula takes a lot of time to get the distributed matrices.
2) Existing social networking services suggest friends to users supported their social graphs, which cannot be the foremost applicable to mirror a user's preferences on friend choice in world.

## IV.PROPOSED SYSTEM

We propose a temporal-topic model to predict user's potential friends. The model initial extracts user's topic distributions from keyword usage patterns of collective messages victimization temporal approach. Then, it calculates user similarities over time supported user's topic distributions. Finally, user's potential interests on others ar foreseen in line with user similarities over totally different periods of your time via temporal functions supported topic model, we tend to conduct friend recommendation to user foreseen scores.

If a user reports others messages with none comments, then system can add "forwarding microblogs" mechanically. Such a denotation doesn't have any result on user'sinterests; thus, we tend to take away it from messages, since reposts messages, however keep the content of the reposted messages, since reposts represents users interests on the connected content.

Suppose that u1 involved "movie" 2 months ago, however currently the most focus is on "landscape," whereas u2 announce microblogsconcerning nature 2 months ago; however currently the posts square measure primarily concerning "film." clearly, interest drifts exist in microblogging. Though u1 and u2 have common interests in movies and nature, as u1's current interest is in nature, u1 might still listen to a nature-related topic, whereas as u2's current interest is in movies, u2 might still concentrate on the moving picture space. During this case, it'll be inappropriate to line a similar weight on all the topics over time for every user to predict another's potential interests.

In distinction, the next weight ought to be allotted to newer topics than those showing a protracted time ago; since newer preferences have bigger influences in predicting users' potential interests than those earlier preferences. Therefore, taking temporal info into thought might improve the accuracy of users' interest prediction.

**Advantages of Proposed System:**

It is effective recommendation system for recommending friends to user.
It takes less time owing to the effectiveness of LDA algorithmic program.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering
(An ISO 3297: 2007 Certified Organization)
Website: www.ijircce.com
Vol. 5, Issue 5, May2017

## V. MATHEMATICAL MODEL

Let S is the Whole System Consist of
S= {U,CD, D, DKE, KC, P, O}.
U = User
U = {u1, u2, …..un}
D = Training Dataset.
CD= Conversation training dataset
D = {d1,d2…..dn}

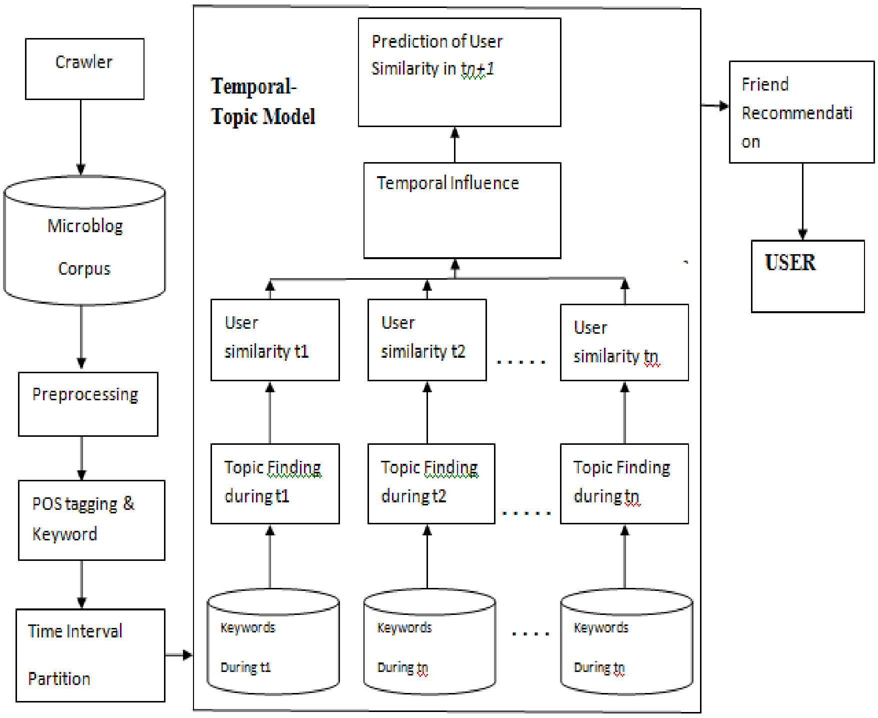
DKE = Diverse keyword extraction
KC = Keyword Clustering
P = probability function.
O = Output.

P = $\sum_{i=1}^{n}(t) * (w)$

Where,

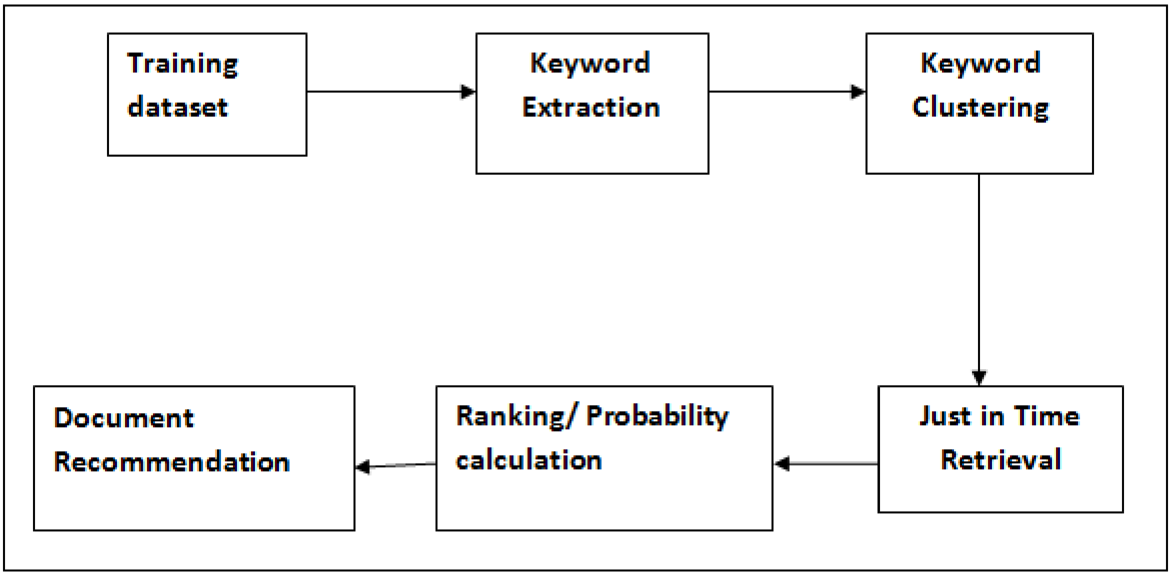
- P is probability function.
- t is total number of topic in the dataset.
- w is the topics/ keyword extracted from dataset.
- n contains the all words in the training document.
- idenotes there exits at least one topic word in the training document.

## VI. SYSTEM ARCHITECTURE



DOI:10.15680/IJIRCCE.2017. 0505234

## VII. WORKING OF FLDA ALGORITHM



### 1) Diverse Keyword Extraction:

The advantage of diverse extractions keyword is that the assurance of the principle subjects of communication fragment is the maximum amount as doable. Moreover, an honest approach of covering bigger subjects, the planned set of rules picks out atiny low amount of key phrases out of each material. That's acceptable for two reasons. This could end in further various implicit queries, for this reason maximising the sort of files retrieved. And, if some words area unit if truth be told, result will generate a primary material within the language fragment, then the set of rules shall recognize atiny low amount of these coaching dataset-full keywords examination with algorithms that forget the foundations of diversity.

### 2) Topic/Keyword Clustering:

Group of topic/keywords (called Clusters) are built by keywords for each main topic of the conversation fragment. One cluster contains similar keywords related to one topic. In our system we used K-means clustering algorithm.

K-means clustering is also known as the centroid based clustering. K-means clustering aims to divide and observations into k clusters in which each observation belongs to the cluster with the closest mean, serving as a example of the cluster. This algorithm randomly selects k points as initial cluster centers. Each point from the dataset is assigned to the closest cluster and each cluster center is then recomputed as the average of points in that cluster. This is repeated until the clusters are formed.

### 2.1 Steps of K-means Algorithm:

K-means clustering algorithm aims to do the partition of 'n' observations into k sets $\{S_1, S_2,\ldots..,S_k\}$ where $k \leq n$ from the given set of observations $(x_1, x_2,\ldots\ldots x_n)$. The steps followed by k-means algorithm are as below:

i. The algorithm randomly selects k points as initial cluster centres.

ii. Each point from the dataset is assigned to the closest cluster based upon the Euclidean distance among each point and each cluster centre.

iii. Each cluster centre is then recomputed as the average of points in that cluster.

iv. Step ii and iii are repeated until the clusters are formed.

**3) Result analysis:**

The result analysis is calculated by following formula (ref from base paper):

mMap = $\frac{\sum_{i=0}^{N_u} MAP\,(i)}{N_u}$

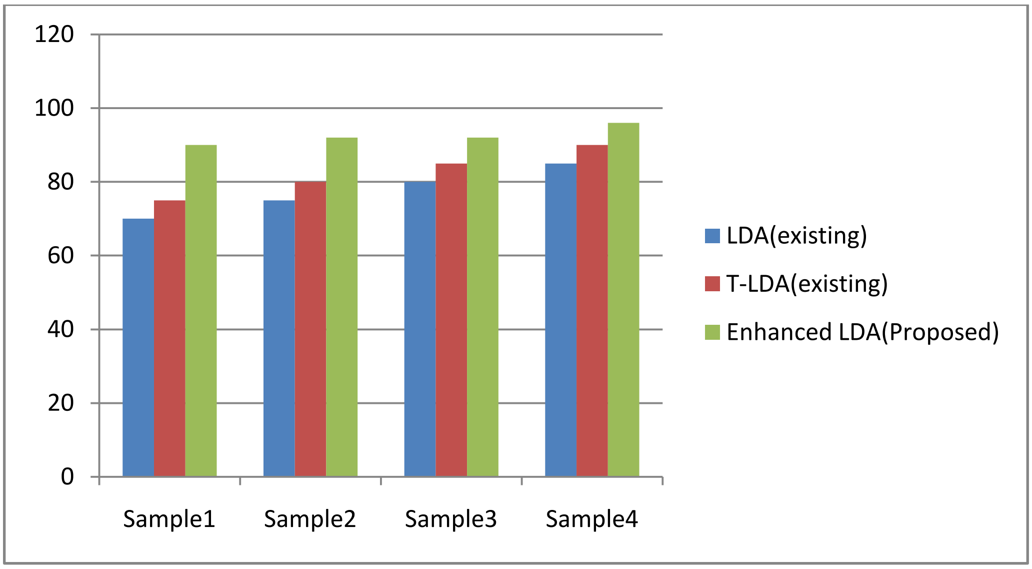
Where,

- MAP($i$) represents the MAP value for the $i$th user.
- $N_u$be the number of user/topics/keywords.
- mMAP is metric of mean average precision.

## VIII. RESULTS

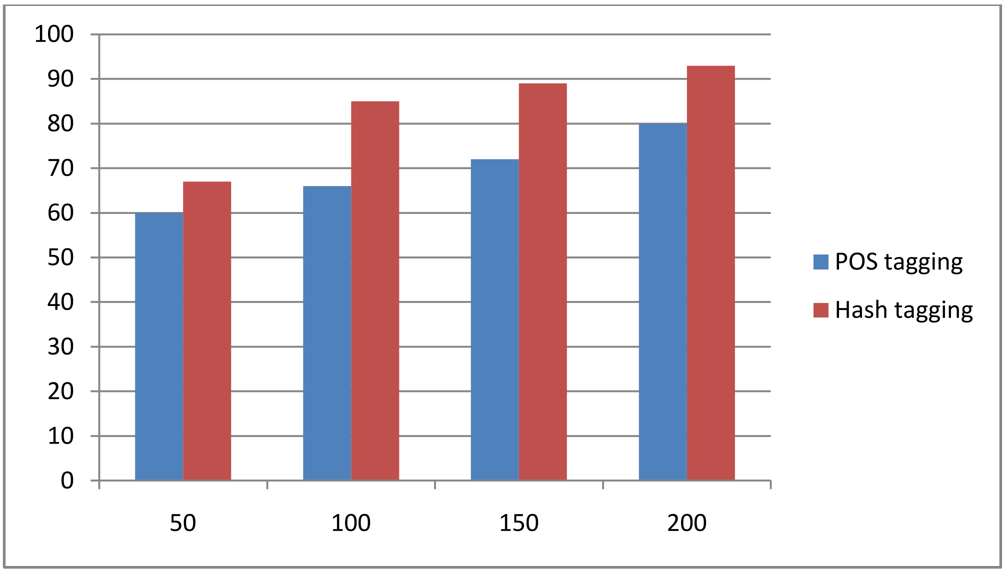
### 1) Efficiency of Proposed System Against Various Existing Systems:

| Samples | LDA(existing)(%) | T-LDA(existing) (%) | Enhanced LDA (Proposed) (%) |
|---|---|---|---|
| Sample1 | 70 | 75 | 90 |
| Sample2 | 75 | 80 | 92 |
| Sample3 | 80 | 85 | 92 |
| Sample4 | 85 | 90 | 96 |



### 2) Efficiency of postagging v/s Hash Tagging:

| Samples | POS tagging(existing) | Hash tagging(proposed) |
|---|---|---|
| 50 | 60 | 67 |
| 100 | 66 | 85 |
| 150 | 72 | 89 |
| 200 | 80 | 93 |



## IX. CONCLUSION

We have planned a disciple recommendation system formula supported a temporal interest evolution model and social tag prediction. The interest evolution model was wont to optimize the immeasurable tags for every microblog. Community discovery and maximization of social tag vote were enforced to predict the tag candidates for a target user.

## REFERENCES

[1] F.-Y. Wang, K. M. Carley, D. Zeng, and W. Mao, “Social computing: From social informatics to social intelligence,” *IEEE Intell. Syst.*, vol. 22, no. 2, pp. 79–83, Mar./Apr. 2007.

[2] C. L. P. Chen and C. Y. Zhang, “Data-intensive applications, challenges, techniques and technologies: A survey on big data,” *Inf. Sci.*, vol. 275, pp. 314–347, Aug. 2014.

[3] M. Moricz, Y. Dosbayev, and M. Berlyant, “PYMK: Friend recommendation at myspace,” in *Proc. ACM SIGMOD Int. Conf. Manage. Data.*,Indianapolis, IN, USA, 2010, pp. 999–1002. D. Ramage, S. Dumais, and D. Liebling, “Characterizing microblogs with topic models,” in *Proc. Int. AAAI Conf. Weblogs Soc. Media*, Menlo Park, CA, USA, 2010, pp. 130–137.

[4] H. Kwak, C. Lee, H. Park, and S. Moon, “What is Twitter, a social network or a news media?” in *Proc. 19th Int. Conf. World Wide Web*, Raleigh, NC, USA, 2010, pp. 591–600.

[5] D. Zhao and M. B. Rosson, “How and why people Twitter: The role that micro-blogging plays in informal communication at work,” in *Proc.ACM Int. Conf. Support. Group Work*, Sanibel Island, FL, USA, 2009,
pp. 243–252.

[6] A. Java, X. Song, T. Finin, and B. Tseng, “Why we Twitter: Understanding microblogging usage and communities,” in *Proc. 9thWebKDD 1st SNA-KDD Workshop Web Min. Soc. Netw. Anal.*, San Jose, CA, USA, 2007, pp. 56–65.

[7] W. X. Zhao *et al.*, “Comparing Twitter and traditional media using topic models,” in *Advances in Information Retrieval*. Berlin, Germany: Springer, 2011, pp. 338–349.

[8] J. Lehmann, B. Goncalves, J. J. Ramasco, and C. Cattuto, “Dynamical classes of collective attention in Twitter,” in *Proc. 21st Int. Conf.World Wide Web*, Lyon, France, 2012, pp. 251–260.

[9] A. Agarwal, V. Xie, I. Vovsha, O. Rambow, and R. Passonneau, “Sentiment analysis of Twitter data,” in *Proc. Workshop Lang.Soc. Media Assoc. Comput. Linguist.*, Stroudsburg, PA, USA, 2011, pp. 30–38.

[10] A. Pak and P. Paroubek, “Twitter as a corpus for sentiment analysis and opinion mining,” in *Proc. Lang. Resour. Eval. Conf. (LREC)*, Valletta, Malta, 2010, pp. 1320–1326.

[11] Armentano, M.G., Godoy, D., Amandi, A.A., 2013. Followee recommendation based on text analysis of microblogging activity.*Inform. Syst.*, 38(8):1116-1127. [doi:10.1016/j.is.2013.05.009]

[12] Balabanović, M., Shoham, Y., 1997. Fab: content-based, collaborative recommendation. *Commun. ACM*, 40(3):66-72. [doi:10.1145/245108.245124]

[13] Breese, J.S., Heckerman, D., Kadie, C., 1998. Empirical analysis of predictive algorithms for collaborative filtering.Proc. 14th Conf. on Uncertainty in Artificial Intelligence, p.43-52.

[14] Cataldi, M., di Caro, L., Schifanella, C., 2010. Emerging topic detection on Twitter based on temporal and social terms evaluation. Proc. 10th Int. Workshop on Multimedia Data Mining, Article 4. [doi:10.1145/1814245.1814249]

[15] Chen, K., Chen, T., Zheng, G., *et al.*, 2012. Collaborative personalized tweet recommendation. Proc. 35th Int. ACM SIGIR Conf. on Research and Development in Information Retrieval, p.661-670. [doi:10.1145/2348283. 2348372]

[16] Chi, C., Liao, Q., Pan, Y., *et al.*, 2011. Smarter social collaboration at IBM research.Proc. ACM Conf. on Computer Supported Cooperative Work, p.159-166. [doi:10.1145/ 1958824.1958848]

[17] Deng, A.L., Zhu, Y.Y., Shi, B., 2003. A collaborative filtering recommendation algorithm based on item rating prediction. *J. Softw.*, 14(9):1621-1628 (in Chinese).

[18] F.-Y. Wang, “Toward a paradigm shift in social computing: The ACP approach,” *IEEE Intell. Syst.*, vol. 22, no. 5, pp. 65–67, Sep./Oct. 2007.