# Survey Paper on LDA Hash Tagging Approach for Friend Recommendation in Microblogging System

**Sumit P. Mirase[1] Prof. N. P. Kulkarni[2]**
[1]Student [2]Assistant Professor
[1,2]Department of Information Technology
[1,2]Smt. Kashibai Navale College of Engineering, Pune, India

*Abstract*— Because of the developing ubiquity and small frame the Microblogging is turning into people's most attractive choice for seeking the information and expressing opinions. Messages got by a user mainly rely on whom user follows. Therefore, recommending user with related interest may enhance the experience quality for information receiving. Since messages posted by Microblogging users reflect their hobbies or interest and the essential keywords in the messages show their primary 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 required, they change as time proceeds by. In light of such instincts, we proposed a temporal-topic to analyze user's possible behavior's and predict their potential friends in Microblogging. The model takes into users' latent preferences by extracting keywords from aggregated messages over a stretch of time using a topic model, and after that, the effect of time is considered to deal interest.

***Key words:*** Microblogging, Temporal, Latent Dirichlet allocation, Semantic enrichment

## I. INTRODUCTION

Microblogging has become a convenient way to Internet surfers and average users to communicate with their friends and family members, or to express intimate emotions or feelings. Using a microblog also has gradually become a habit for a massive amount of users, which leads to an exponential explosion of information in the virtual microblog society on the Internet, making retrieving and identifying needed microblog or related information strict. Therefore, more and more microblog services are developing different engines dedicated to recommending user-specific information.

Early researchers mainly focused on the features of Microblogging and social network analysis. Recently, there has been an increasing interest in the field of information retrievals, such as event detection and tracking, identification of influential people, sentiment analysis, and personalized recommendations.

Traditional recommendation systems can primarily classify into three categories: CF-based, content-based, and hybrid recommendation systems Probabilistic topic models have been shown to be the reliable tools for identifying latent text patterns in the content. Latent Dirichlet allocation (LDA) delivers the capacity of concluding the topic distributions so that the model can be used to generate hidden documents as well. LDA has also been applied to various works on Twitter to demonstrate its usefulness. Users' interests are not inactive; contrarily, their interests may change as time passes. Since the real-time and brevity features of Microblogging lead to frequent updates of microblog, users' interests are more extensive and changeable over time.

## II. MOTIVATION

Microblogging e.g. Twitter as a new kind of online communication in which users discuss their routine lives, share information by short columns or publish scenes, have become the most preferred social networking services today, makes it possibly a huge knowledge base attracting increasing attention of researchers in the field of knowledge discovery and data mining.

Messages received by a user mainly depend on whom the user follows. Thus, to recommend users with similar interests may improve user's expertise for information they desire to gain. Users regularly post microblogs to record daily life and express opinions. Therefore, posts published by users, to some extent, reflect their interests. By mining user's social behaviors and dynamics, we may help them find friends with similar interests, which may improve the users' experience, social interactions, and gain more business value for corporations.

## III. RELATED WORK

### A. *Characterizing Microblogs with Topic Models.[1]*

As microblogging rises in popularity, services like Twitter are coming to support information gathering needs above and exceeding their regular roles as social networks. But most users' communication with Twitter is still primarily focused on their social graphs, forcing the often improper conflation of "people I follow" with "stuff I want to read." We characterize some information needs that the current Twitter interface fails to support, and maintain for excellent representations of content for solving these challenges. We offer scalable implementation of a partly managed learning model (Labeled LDA) that outlines the content of the Twitter feed inside dimensions. These dimensions correspond roughly to style, status, substance, and social characteristics of posts. We characterize users and tweets using this model, and existing results on two information consumption oriented tasks.

### B. *What is Twitter, a Social Network or a News Media? [2]*

The goal of this paper is to study the topological characteristics of Twitter and its power as a new medium of information sharing.

Ranking by retweet differs from the prior two rankings indicates a gap in influence inferred from plenty of followers and that from the popularity of one's tweets. We have examined the tweets of top trending topics and reported on their temporal behavior and user participation. We have listed the trending topics based on the active the period and the tweets and show that the majority (over 85%)

 

of points are headline news or persistent news in nature. Once retweeted, a tweet gets retweeted nearly instantly on next hops, signifying fast dissemination of information after the 1st retweet.

*C. How and Why People Twitter: The Role that Micro-blogging plays in Informal Communication at Work [3]*

Micro-blogs, a comparatively new phenomenon, provide a new communication channel for people to distribute information that they likely would not share unless using existing channels (e.g., email, phone, IM, or weblogs). Micro-blogging has become popular quite quickly, raising its potential for serving as a new informal communication medium at work, providing a kind of impacts on collaborative work (e.g., enhancing information sharing, building common ground, and sustaining a feeling of connectedness among colleagues). This exploratory research project is pointed at obtaining an in-detail understanding of how and why people use

Twitter – a famous micro-blogging tool - and exploring microblogs potential impacts on informal communication at work.

*D. Detecting Popular Topics in Micro-blogging Based on a User Interest-Based Model [4]*

Twitter as a new form of social media can potentially contain much useful information, but content analysis on Twitter has not been well thought. In particular, it is not clear whether as an information source Twitter can be regarded as faster news feed that covers mostly the similar information as traditional news media. In This paper we empirically compare the content of Twitter with a traditional news medium, New York Times, using unsupervised topic modelling. We use a Twitter-LDA model to discover topics from a representative sample of the entire Twitter. We then use text mining techniques to compare these Twitter topics with topics from New York Times, taking into consideration topic categories and types. We also study the relation between the proportions of opinionated tweets and retweet and topic categories and types. Our comparisons show interesting and useful for downstream IR or DM applications.

*E. Detecting Popular Topics in Micro-blogging Based on a User Interest-Based Model [5]*

The rapidly increasing popularity of micro-blogging has made it an important information seeking channel. By detecting recent popular topics from micro- logging, we have opportunities to gain insights into internet hotspots. Two primary factors determine a topic's popularity. One is how a issue is considered by users, and the other is how much influence those users have since points shown in the primary users' posts are more likely to attract others' attention. However, existing methods interpret a topic's popularity with only the number of keywords related to it, which neglect the value of the user influence to information diffusion in micro-blogging. In this paper, drawing upon the Cognitive Authority Theory and Social Network Theory, we introduce a novel model that recognizes the most popular topics in micro-blogging with a user interest-based method. The proposed model first creates a topic graph according to users' concerns and their following relationship and calculates the topics' popularity with a link-based ranking algorithm. The familiar topics discovered by the method can reflect the relationship between the topics in the posts and users' interests of influential users can be highlighted. Experimental conclusions on the data of Twitter, an outstanding and feature-rich micro-blogging service, show that the proposed method is effective in popular topic development.

*F. Twitter Rank: Finding Topic-sensitive Influential Twitters [6]*

This paper focuses on the problem of knowing influential users of micro-blogging services. Twitter, one of the several well-known micro-blogging services, employs a social-networking model called "following," in which each user can pick who she wants to "follow" to receive tweets from outwardly requiring the latter to give permission first. In a dataset developed for this study, it is observed that (1) 72.4% of the users in Twitter follow more than 80% of their followers, and (2) 80.5% of the users have 80% of users they are following follow them back. Our study reveals that the presence of "reciprocity" can be explained by aspect of homophile. Based on this finding, Twitter Rank, an extension of Page Rank algorithm, is proposed to measure the influence of users in Twitter. Twitter Rank measures the influence taking both the topical similarity between users and the link structure into account. Experimental results show that Twitter Rank outperforms the one Twitter currently uses and other related algorithms, including the original Page Rank and Topic-sensitive Page Rank.

*G. Measuring User Influence in Twitter: The Million Follower Fallacy [7]*

Directed links in social media could serve anything from special friendships to mutual interests, or even a passion for breaking news or celebrity gossip. Such organized links define the flow of information and hence indicate a user's impact on others—a concept that is crucial in sociology and viral marketing. In this paper, using a large amount of data assembled from Twitter, we present an in-depth comparison of three areas of influence: in degree, retweet, and mentions. Based on these measures, we investigate the dynamics of user influence across topics and time. We make several interesting observations. First, popular users who have high in degree are not necessarily influential regarding spawning retweet or mentions. Second, most influential sers can hold meaningful influence over a variety of topics. Third, influence is not gained spontaneously or accidentally, but through collective effort such as limiting tweets to a single topic. We believe that these conclusions provide new insights for viral marketing and suggest that topological measures such as in degree alone exposes very small about the influence of a user.

## IV. DISCUSSION AND COMPARISON OF EXISTING WORK

By above survey, we compare some mechanism used for friend recommendation.

 

| Paper Name | Platform used and Languages Supported | Advantages | Limitations |
|---|---|---|---|
| Characterizing Microblogs with Topic Models[1] | Platform-Java doc in JDK 1.7,1.6 Microsoft Visual Basic 6.0 and Microsoft Access 97 Languages supported C#, Java, Smalltalk | Shows scalable implementation of partly supervised learning model (Labelled LDA) that maps the content of the Twitter feed into dimensions. These dimensions correspond approximately to substance, style, status, and social features of posts. We characterize users and tweets using this model, and present results on two information consumption oriented tasks. | This does not work takes richer content-based analysis of Twitter, we understand there is a bright future for such models on microblogs moving forward. |
| What is Twitter, a Social Network or a News Media?[2] | Platform-Java doc in JDK 1.7,1.6 Microsoft Visual Basic 6.0 and Microsoft Access 97 Languages supported C#, Java | Study the topological qualities of Twitter and its power as the advanced medium of information sharing. | Twitter with its free API to crawl, one-sided nature of the relationship, and the retweet tool to relay information offers an unprecedented opportunity for sociologists, linguists, computer scientists, and physicists to study human performance. Our work is the first step towards exploring the numerous potentials of this new platform. |
| How and Why People Twitter: The Role that Micro-blogging Plays in Informal Communication at Work[3] | Platform-Microsoft Visual Basic 6.0 and Microsoft Access 97 Languages supported C#, .Net. | Understand how and why people use Twitter – a popular micro-blogging tool - and exploring microblogs potential impacts on informal communication at work. Provide a distinct communication channel for people to broadcast information that they likely would not share | The cost of sharing and promoting more frequent updates in real-time, as well as making it easier for users to browse and monitor a large amount of information updates. |
| Comparing Twitter and Traditional Media using Topic Models[4] | Platform-Microsoft Visual Basic 6.0 and Microsoft Access 97 Languages supported C#,Net | Compared the content of Twitter with a normal traditional news medium, New York Times, concentrating on the differences betweenthese two. We formed a new Twitter-LDA model that is designed for short tweets and proved its effectiveness compared with existing models. Our analysis of the topical variations between Twitter and traditional news media. Our empirical comparison confirmed some previous views and also revealed some new findings. | Not study how to summarize and visualize Twitter content In a systematic way. Our method of associating tweets with distinct categories and types may also help visualization of Twitter content. |
| Detecting Popular Topics in Micro-blogging Based on a User Interest-Based Model[5] | Platform-Java doc in JDK 1.7,1.6 Microsoft Visual Basic 6.0 and Microsoft Access 97 Languages | Micro-blogging is essentially used to note what is happeningaround the world and to strengthen communication between users in a social network | Not detecting the evolution of favorite topics which is due to the changing of users' interest, we plan to introduce a time-sensitive model for topic listing not only in Micro-Blogging but also in other social networks. Domain sensitive topic ranking based on users' interest |

 

| | | | |
|---|---|---|---|
| | supported<br>C#, Java | | |
| Twitter Rank: Finding Topic-sensitive Influential Twitters[6] | Platform-Microsoft Visual Basic 6.0 and Microsoft Access 97<br>Languages supported<br>C#, Net | Identify influential users of micro-blogging services. *Twitter*, one of the most notable micro-blogging services. | Twitter is a platform for free and opens conversations among twitters. An incremental approach to topic distillation in *Twitter* is still a topic deserves further study. |
| Measuring User Influence in Twitter: The Million Follower Fallacy[7] | Platform-Java doc in JDK 1.7,1.6 Microsoft Visual Basic 6.0 and Microsoft Access 97<br>Languages supported<br>C#, Java | Directed links in social media could serve anything from private friendships to mutual interests, or evena feeling for breaking news or celebrity gossip. | Not gained automatically but through concerted effort. In order to gain and maintain influence, users require to keep great personal involvement. This could mean that influential users are further predictable than suggested by theory (Watts 2007), shedding light on how to identify emerging influential users. |

Table 1: Summery and comparison of existing work

## V. PROPOSED WORK

### A. *Modules:*

1) Preprocessing.
2) POS and Keyword Extraction.
3) Time Interval Partition.
4) Topic Finding.
5) User Similarity Calculation.
6) Temporal Influence.
7) Friend Recommendation.

## VI. MODULE DESCRIPTION

### A. *Preprocessing:*

In some systems like, Sina Weibo, if a user reposts others' messages without any comments, the system will add, "forwarding microblogs" automatically. Such a definition does not have any impact on users' interests; therefore, we dismiss it from messages, but retain the content of the reposted messages, since reposts represent users' interests on the related content. Additionally, we remove URLs and other no texts from microblogs.

### B. *Hash Tagging and Keyword Extraction:*

In this module, we perform word segmentation and Hash tagging for messages. We apply word segmentation platform to pre-process the corpus. The segmentation platform proposes a word segmentation approach based on the integration of human intelligence, big data, and machine learning. Based on Hash tagging, we extract nouns, abbreviations, idioms, and academic vocabularies as meaningful notional words that form keywords for further analysis.

### C. *Time Interval Partition:*

Users' interests vary as time goes by, which reveals and users' microblogs may focus on different topics at different periods of time. Therefore, users' dynamically varying interests can be expressed as a sequence of keyword combinations in microblogs at various time intervals, i.e., M = M1 U M2. U Mn. Each Mt denotes a temporal user-keyword matrix at the tth time interval, where $M_t \in R^{Nu \times Nw}$ And Nu & Nw are the numbers of users and keywords, respectively. Each row of Mt includes the word counts at the tth time interval for a particular user, whereas each column of Mt contains the counts by different users for a certain word at the tth time interval.

### D. *Topic Finding:*

Only keywords are not sufficient for determining users' interests. As the existence of synonymy, it needs to find the hidden topics from the keyword usage patterns. Since the aim is to find topics that each microblogging user is interested in rather than topics that each microblog is about, we handle the microblogs published by an individual user at the tth period as a big document. Then, each row of sub-collection Mt is treated as a bag-of-words document that essentially corresponds to a user. To find user materialistic topics in Mt, or to find temporal topics of every document in Mt, we apply the LDA model. Each user is associated with a mixture of different topics, and each topic is represented by a probabilistic distribution over keywords. Formally, each of a collection of Nu users is associated with a multinomial distribution over T topics, which is denoted as θu(t) at time t. Each topic is associated with a multinomial distribution over keywords, denoted as φz(t). θu(t) and φz(t) have Dirichlet prior with hyper-parameters αt and βt, respectively. For each keyword of user u, a topic zt is sampled from the multinomial distribution θu(t) associated

 

with user u at time t, and a keyword wt from the multinomial distribution φz(t) correlated with topic zt is sampled consequently. This generative process is repeated Nuw (t) times to form user u's collection of keywords.

### *E. User Similarity Calculation*

After row normalizing θ (t) to θ (t), the ith row of matrix θ(t) presents a linear additive mixture of factors to indicate user i's interests over T topics at the tth time interval. The larger weight user i am assigned to a factor, the more interest user i have in the relevant topic. It has been demonstrated in that microblogger follows a friend because he is interested in some topics the friend is publishing. Hence, for friend recommendations, we aim to find users' topic similarity based on the normalized user-topic distribution θ(t).

### *F. Temporal Influence*

In this module, we desire to utilize users' sequential topical similarity matrices {S1, S2, . . . , Sn} to predict users' potential interests shortly. Generally speaking, users' historical favorites may impact his future interests, and more recent interests may have the stronger impact on the future preference prediction than earlier interests. To imitate the influence of historical behaviors, we apply the exponential decay function, which has been proved to be an effective function to measure interest drifts.

### *G. Friend Recommendation*

Finally, we utilize the exponential decay function with kernel parameter γ to predict users' potential interests on others at time t.

$$P_n = \sum_{t=1}^{n-1} \exp\left(-\frac{n-t}{\gamma}\right) S_t$$

Where Pn is the probability matrix of potential interests among users at time n. A higher score means that the two users have a greater association and that they may have the higher likelihood of becoming a friend. Finally, users are classified by the score and those with higher scores are recommended to the target user.

## VII. CONCLUSION AND FUTURE WORK

In this project, we propose a temporal-topic model for friend recommendations microblogging systems. The model first discovers users' latent preferences i.e. hash tags during different time intervals based on keywords extracted from the summed microblogs through a topic model. Then, it calculates user identities in each time interval based on sequence of users' interests with the timeline. Based on the model, we conducted friend recommendations and the experimental results showed that our model is effective.

For future work, we plan to conduct our experiments on users who have less friends and followers to prove if our model is useful for the cold-start problem of personalized recommendations. We also aim to unearth other factors to enhance the performance of the proposed model, such as social relationships among users (i.e., followers, Followee), the sentiment of microblogs, users' location information, etc. We also plan to examine other state-of-the-art models with temporal evolvement and compare the performances of different methods on friend recommendations. Other datasets such as Twitter will be tested for the usefulness and effectiveness of the model.

## REFERENCES

[1] Daniel Ramage, Susan Dumais, Dan Liebling,"Characterizing Microblogs with Topic Models", Proceedings of the Fourth International AAAI Conference on Weblogs and Social Media,2010, pp. 130-138

[2] Haewoon Kwak, Changhyun Lee, Hosung Park, and Sue Moon What are Twitter, a Social Network or a News Media? WWW 2010, April 26–30, 2010, Raleigh, North Carolina, USA. ACM 978-1-60558-799-8/10/04

[3] Dejin Zhao , Mary Beth Rosson How andWhy People Twitter: The Role that Micro-blogging Plays in Informal Communication at Work, *GROUP'04*, May 10–13, 2009, Sanibel Island, Florida, USA. Copyright 2009 ACM 978-1-60558-500-0/09/05

[4] Zhao, Wayne X., Jiang Jing, Wng Jianshu, He Jing, Lim Ee-Peng, Yan Hongfei and Li Xiaoming. 2011. Comparing Twitter and Traditional Media Using Topic Models. In Advances in Information Retrieval: 33rd European Conference on IR Research, ECIR 2011, Dublin, Ireland, April 18-21, 2011.

[5] Shuangyong Song, Qiudan Li, Xiaolong Zheng "Detecting Popular Topics in Micro-blogging Based on a User Interest-Based Model", WCCI 2012 IEEE World Congress on Computational Intelligence, - Brisbane, Australia June, 10-15, 2012

[6] WENG, Jianshu; LIM, Ee Peng; JIANG, Jing; and He, Qi. Twitterrank: Finding Topic-Sensitive Influential Twitterers. (2010). ACM International Conference on Web Search and Data Mining (WSDM 2010). , 261..

[7] Meeyoung Cha, Hamed Haddadiy, Fabr´ıcio Benevenutoz, and Krishna P. Gummadi_ "Measuring User Influence in Twitter: The Million Follower Fallacy".

 