Latent Dirichlet Allocation Hash Tagging Approach for Friend Recommendation in Microblogging System

Sumit P. Mirase[1] Prof. N. P. Kulkarni[2]
[1,2]Department of Information Technology
[1,2]Smt. Kashibai Navale Collage of Engineering, Vadgaon (BK), Pune, India

Abstract— Microblogging is becoming people's most attractive choice for getting the information and expressing opinions because of the developing universality and small frame. 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 important keywords in the messages show their primary focus to a huge extent, we can find users' interest by investigating the user generated contents. Besides, user's hobbies, interest are not static; despite what might be required, they change as time changes. In light of such instincts, we proposed a LDA model in microblogging system for friend recommendation to analyze user's possible behavior's and predict their potential friends in Microblogging. The model takes into users' potential preferences by extracting keywords from aggregated messages over a period 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 makes it possibly a huge knowledge base attracting increasing attention of researchers in the field of knowledge discovery and data mining. Consider the example of Twitter in which users discuss their routine lives, share information by short columns or publish scenes, have become the most preferred social networking services today, 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 behaviours 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.

The goal of the proposed system is to predict users' potential interests using temporal latent semantic analysis (LDA model). Also to enhance the LDA model using time interval partition which consider change in users' likes and interests. Based on this user's likes and interest others should get friend recommendation.

The remaining paper is organized as follows. Section II covers related work, Section III contains implementation details, Section IV covers Results and Discussion and section V contains Conclusion and Future scope.

II. RELATED WORK

The tool presented by Daniel Ramage, Susan Dumais, Dan Liebling [1] offer scalable implementation of a partly managed learning model i.e. LDA that outlines the content of the Twitter feed inside dimensions. This characterizes users and tweets using this model, and existing results on two information consumption oriented tasks. Haewoon Kwak, Changhyun Lee, Hosung Park, and Sue Moon [2] states to study the topological characteristics of Twitter and its power as a new medium of information sharing. Dejin Zhao, Mary Beth Rosson [3] 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. Zhao, Wayne X., Jiang Jing, Wng Jianshu, He Jing, Lim Ee-Peng, Yan Hongfei and Li Xiaoming [4] use a Twitter-LDA model to discover topics from a representative sample of the entire Twitter. Also use text mining techniques to compare these Twitter topics with topics from New York Times, taking into consideration topic categories and types. Shuangyong Song, Qiudan Li, Xiaolong Zheng [5] 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. WENG, Jianshu; LIM, Ee Peng; JIANG, Jing; and He, Qi [6] focuses on the problem of knowing influential users of micro-blogging services and 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. Meeyoung Cha, Hamed Haddadiy, Fabr´ıcio Benevenutoz, Krishna P. Gummadi [7] investigates the dynamics of user influence across topics and time. This makes several interesting observations. First, popular users who have high in degree are not necessarily influential regarding spawning retweet or mentions. Second, most influential 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.

III. IMPLEMENTATION DETAILS

In this paper, we design a tool for friend recommendation based on user's likes and interests.

Fig. 1 shows the system architecture of LDA hash tagging system.

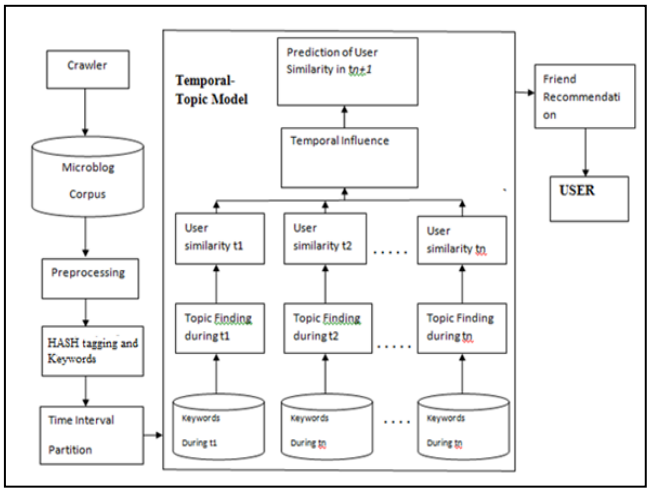


Fig. 1: System Architecture of LDA hash tagging system

The proposed tool contains the following modules:

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 [1].

Each Mt denotes a temporal user-keyword matrix at the tth time interval, where $M_t \in R^{Nu \times Nw}$ [1] 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) provides a linear additive combination of factors to indicate user i's interests over T topics at the tth time interval. The higher weight user i is assigned to a factor, the more interest user i has 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. Therefore, 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 favourites may influence 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 behaviours, we apply the exponential decay function, which has been proved to be an effective function to measure interest drifts.

G. Friend Recommendation

Finally, users are classified by the score and those with higher scores are recommended to the target user.

1) Mathematical Model:

Let W be the whole system which consists:

W= {U, W, Nu, Nw, t, T, S, St,I,M}.

Where,

1) U is the set of user.
   U= {U1, U2 …Un}
2) W is the set of keywords.
   W= {W1, W2 …Wn}
3) Nu is the set of total number of user.
   Nu= {Nu1, Nu2 …Nun}
4) Nw is the set of total number of keywords.
   Nw= {Nw1, Nw2 …Nwn}
5) t is the time interval.
   T= {t1,t2,…tn}
6) T is the set of number of topics.
   T= {T1, T2,…Tn}

7) S is the set of similarity matrix.
S= {S1, S2,…Sn}
8) St is the set of users topical similarity matrix at time t.
St= {St1, St2,…Stn}
9) I is set of the number of iterations in LDA model.
I= {I1, I2,…In}
10) M is the set of keyword matrix.
M= {M1, M2,…Mn}

a) Step 1: Login the number of users
U= {U1, U2,…Un}

b) Step 2: Process the number of keywords from the users
W= {W1, W2,…Wn}

The keywords i.e. likes may change on time basis i.e.
M = M1 UM2, ... ,UMn [1]
Where Mt denotes a temporal user-keyword matrix at the tth time interval

c) Step 3: Generate the similarity matrix
S= {S1,S2,…Sn}
Based on the likes similarity matrix is created:
Djs(i,j)=1/2(DKL(Θ(t)i).||H(t))+DKL(Θ'(t)j.(i(t))) [1]
Where,
Djs is the multinomial distribution topic specific to the user u at time t
Dkl is multinomial distribution words specific to the topic z at time t
Then based on similarity matrix friend recommendation is done.

$$f(t) = \exp(-\frac{n-t}{\gamma}(t\epsilon\{1,2,\dots,n-1\}, \gamma > 0)$$

Where,
f(t) is friend recommendation.
n is the total number of time interwals
t is the time interval.
γ is the kernel parameter in the exponential decay function.
Output: Friend Recommendation

IV. RESULTS AND DISCUSSION

The Efficiency of proposed system against various existing systems:

We compare our proposed LDA system with existing systems as shown in Table 1. The efficiency of proposed system is far better than the existing once.

Suppose we consider 5 samples w.r.t. LDA system, T-LDA system and our proposed system, as shown in below Fig. 1 we come to the conclusion that our proposed system is more efficient.

Samples	LDA(existing) (%)	T-LDA(existing) (%)	Enhanced LDA (Proposed) (%)
Sample 1	70	75	90
Sample 2	75	80	92
Sample 3	80	85	92
Sample 4	85	90	96

Table 1: Comparison between Efficiency of proposed system against various existing systems

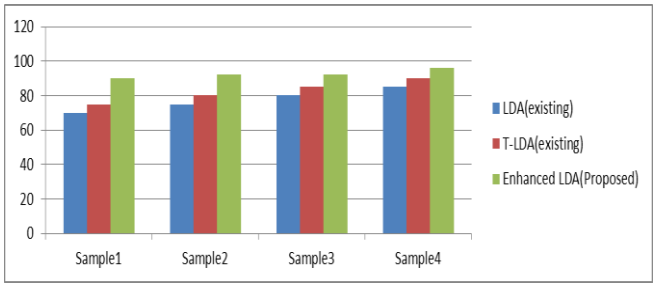


Fig. 1: Comparison between Efficiency of proposed system against various existing systems

Efficiency of pos tagging v/s hash tagging:

We compare existing pos tagging algorithm with proposed hash tagging algorithm as shown in Table 2. The efficiency of proposed hash tagging algorithm is far better than pos tagging algorithm.

Suppose we consider 5 samples w.r.t. POS tagging system and our proposed hash tagging system, as shown in below Fig. 2 we come to the conclusion that our proposed hash tagging system is more efficient.

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

Table 2: Comparison between Efficiency of pos tagging v/s hash tagging

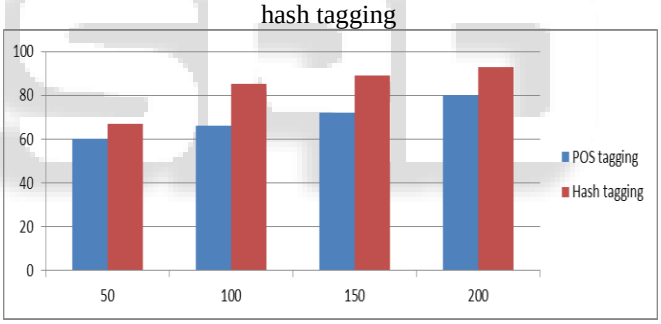
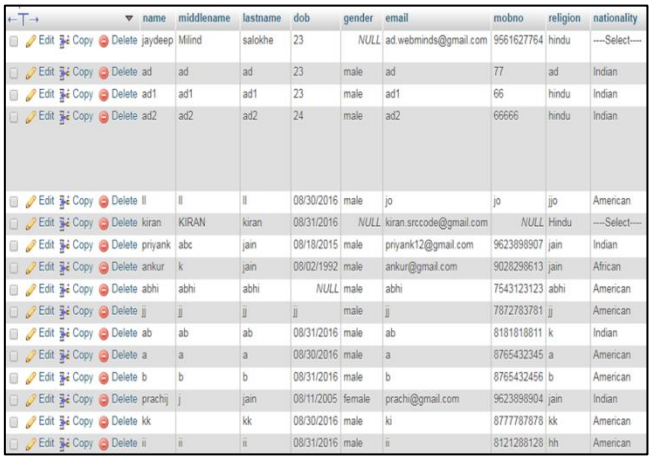


Fig. 2: Comparison between Efficiency of pos tagging v/s hash tagging

In our proposed system there are two types of users Celebrities and Members. Based on similarities between users likes and hash tags users and celebrities get recommend to each other.



	name	middlename	lastname	dob	gender	email	mobno	religion	nationality
Edit Copy Delete	jaydeep	Milind	salokhe	23	NULL	ad.webminds@gmail.com	9561627764	hindu	----Select----
Edit Copy Delete	ad	ad	ad	23	male	ad	77	ad	Indian
Edit Copy Delete	ad1	ad1	ad1	23	male	ad1	66	hindu	Indian
Edit Copy Delete	ad2	ad2	ad2	24	male	ad2	66666	hindu	Indian
Edit Copy Delete	ll	ll	ll	08/30/2016	male	jo	jo	jjo	American
Edit Copy Delete	kiran	KIRAN	kiran	08/31/2016	NULL	kiran.srccode@gmail.com	NULL	Hindu	----Select----
Edit Copy Delete	priyank	abc	jain	08/18/2015	male	priyank12@gmail.com	9623898907	jain	Indian
Edit Copy Delete	ankur	k	jain	08/02/1992	male	ankur@gmail.com	9028298613	jain	African
Edit Copy Delete	abhi	abhi	abhi	NULL	male	abhi	7543123123	abhi	American
Edit Copy Delete	jj	jj	jj	jj	male	jj	7872783781	jj	American
Edit Copy Delete	ab	ab	ab	08/31/2016	male	ab	8181818811	k	Indian
Edit Copy Delete	a	a	a	08/30/2016	male	a	8765432345	a	American
Edit Copy Delete	b	b	b	08/31/2016	male	b	8765432456	b	American
Edit Copy Delete	prachij	j	jain	08/11/2005	female	prach@gmail.com	9623898904	jain	Indian
Edit Copy Delete	kk		kk	08/30/2016	male	ki	8777787878	kk	American
Edit Copy Delete	ii	ii	ii	08/31/2016	male	ii	8121288128	hh	American

Fig. 3: Member database

As shown in Fig. 3 Member database, as the intlike and hash tag fields of users get match with each other, it will directly get recommend to each other.

As shown in below Fig.4 Celebrity or member can edit their profile also write their own blogs. The Home button redirects user to Logged in user's home profile. Blog button allows writing our blog. Edit Celebrity Profile button allows to edit the profile data. About us gives the details of our system and Logout to get out of the session.

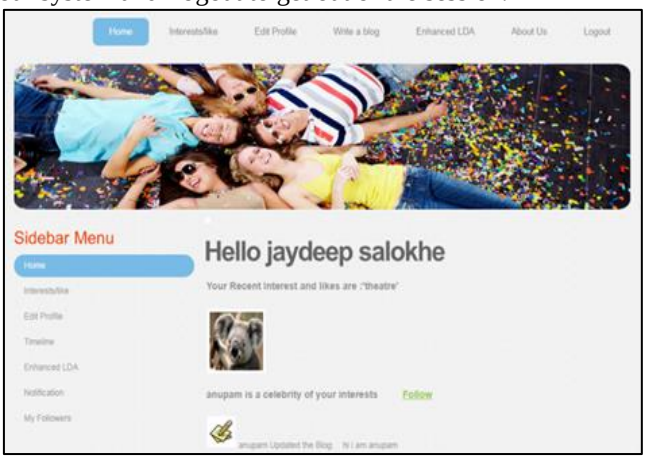


Fig. 4: Member home page GUI

V. Conclusion and Future Scope

In this project, we proposed a LDA based model in microblogging system for friend recommendations. The model first discovers users' potential preferences during different time intervals based on keywords extracted from the aggregated microblogs through a topic model. Then, it calculates user similarities in each time interval based on temporal topic distributions. After that, an exponential decay function is used to measure interest drifts. Finally, users' potential interests on others can be predicted based on the sequence of users' interests along 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 show 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 investigate 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 is 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 and Why 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 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, Krishna P. Gummadi_ "Measuring User Influence in Twitter: The Million Follower Fallacy".

