Survey Paper on Friend Recommendation Method Using Physical and Social Context of Twitter Timeline

Manish Narkhede[1], Pratik Shelar[2], Kiran Sonawane[3], Ujwal Chaudhari[4]

[1,2,3,4] Information Technology,SKN SITS Lonavala,Pune, Maharashtra, India.
Assistant Professor. Shruti Agrawal, Dept. of Information Technology ,SKN SITS Lonavala,Pune, Maharashtra, India.

---***---

Abstract - *Because of the brief frame and developing ubiquity the Micro blogging is turning into people's most interesting choice for seeking the information and expressing opinions. Messages got by a user mainly rely on whom user follows. Therefore, recommending user with comparable interest may enhance the experience quality for information receiving. Since messages posted by Micro blogging 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. In light of such instincts, we proposed a temporal-topic to analyze user's possible behaviors' and predict their potential friends in Micro blogging. The model takes in users' latent preferences by extracting keywords on aggregated messages over a stretch of time by means of a topic model, and after that the effect of time is considered to deal interest.*

Key Words: *Encryption,Descryption,Social Networking*

1.INTRODUCTION

Micro blogging has become a convenient way for Internet surfers and average users to communicate with their friends and family members, or to express intimate emotions or feelings. Using a micro blog also has gradually become a habit for a massive amount of users, which leads to an exponential explosion of information in the virtual micro blog society on the Internet, making retrieving and identifying needed micro blog or related information extremely difficult. Therefore, more and more micro blog services are developing novel engines dedicated to recommending user-specific information.Early researchers mainly focused on the characteristics of Micro blogging and social network analysis. Recently, there has been an increasing interest in the field of information retrieval, such as event detection and tracking, identification of influential people, sentiment analysis, and personalized recommendations.

Traditional recommendation systems can primarily be classified into three categories: CF-based, content-based, and hybrid recommendation systems Probabilistic topic models have been proved to be the powerful tools for identifying latent text patterns in the content. Latent Dirichlet allocation (LDA) achieves the capacity of generalizing the topic distributions so that the model can be

used to generate unseen documents as well. LDA has also been applied to various works on Twitter to demonstrate its usefulness. Users' interests are not static; contrarily, their interests may change as time goes by. Since the real-time and brevity features of Micro blogging lead to frequent updates of micro blog, users' interests are more extensive and changeable over time.

2. LITERATURE SURVEY

2.1 Barbosa and Feng (2010)

The significant effort for sentiment classification on Twitter data is given by Barbosa and Feng. They use polarity predictions from three websites as noisy labels to train a model and use 1000 manually labeled tweets for tuning and another 1000 manually labeled tweets for testing. They however do not mention how they collect their test data. They propose the use of syntax features of tweets like retweet, hashtags, link, punctuation and exclamation marks in conjunction with features like prior polarity of words and POS of words. We extend their approach by using real valued prior polarity, and by combining prior polarity with POS. Our results show that the features that enhance the performance of our classifiers the most are features that combine prior polarity of words with their parts of speech. The tweet syntax features help but only marginally.

2.2 Gamon (2004):

Perform sentiment analysis on feeadback data from Global Support Services survey. One aim of their paper is to analyze the role of linguistic features like POS tags. They perform extensive feature analysis and feature selection and demonstrate that abstract linguistic analysis features contributes to the classifier accuracy. In this paper we perform extensive feature analysis and show that the use of only 100 abstract linguistic features performs as well as a hard unigram baseline.

2.3 Bring & Page:

Bring& Page has introduced the Page Rank algorithm. recomputed a rank vector that provides a priori authority estimates for all of the nodes in a given graph. The node authority is independent of the attributes of each node and such an authority measure only emerges from the topological structure of the graph. In particular, the authority of a node m depends on the number of incoming links and on the authority of the nodes which point to m with forward links. In this paper, a Page Rank based model is proposed to detect the most popular topics in micro-blogging based on users' interest relationship. The model first detects the favorite topics of each user with voting theory, then creates the links between topics with users' interest relationship to construct the 'topic graph' in the whole micro-blogging social network, finally, ranks those topics with Page Rank algorithm to find the most popular ones in micro-blogging

3.EXISTING SYSTEM

In existing traditional recommendation systems, many prediction algorithms, such as the singular value decomposition (SVD) based algorithm, are then conducted directly on these sparse matrices to fill out the missing elements. Considering the cold-start problem, before prediction and recommendation, we optimize the tags of micro blog using the interest evolution model and initialize user preference with the cold start problem by social tag prediction.Most of the present friend suggestions mechanism relies on pre-existing user relationships to pick friend candidates like friend of friend i.e. mutual friends

3.1 Disadvantages of Existing System:

1. The SVD based algorithm takes much time to generate the sparse matrices.

2. Existing social networking services recommend friends to users based on their social graphs, which may not be the most appropriate to reflect a user's preferences on friend selection in real life.

4. PROPOSED SYSTEM

We propose a temporal-topic model to predict users potential friends. The model first extracts user's topic distributions from keyword usage patterns of aggregated messages using temporal approach. Then, it calculates user similarities over time based on users topic distributions. Finally, users potential interests on others are predicted according to user similarities over different periods of time via temporal functions based on topic model, we conduct friend recommendation to user predicted scores.If a user reports others messages without any comments, then system will add "forwarding microblogs" automatically. Such a denotation does not have any effect on users intersts; therefore, we remove it from messages, since reposts messages, but keep the content of the reposted messages, since reposts reporesnts users interests on the related content.

5. SYSTEM ARCHITECTURE

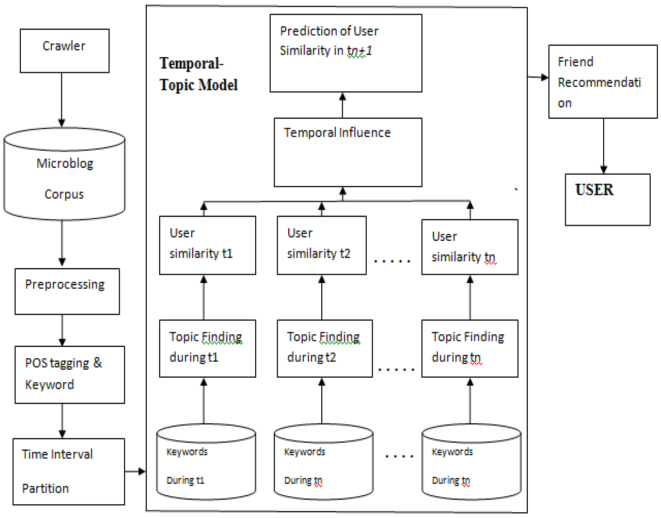

Fig. : System Architecture

6. MATHEMATICAL MODEL

Let W be the whole system which consists:

W= {U, W, Nu , Nw, W, t, T , β, α, θ, γ, δ, , I, n, w, S, M}.Where,

1. U is the set of user.
2. W is the set of keywords.
3. Nu is the set of total number of user.

4. Nw is the set of total number of keywords.
5. t is the time interval.
6. N_w^u (t) is the total number of keywords of user u at time t.
7. T is the set of number of topics.
8. n is the total number of time interwals.
9. β is the Dirichlet prior for user.
10. α is the Dirichlet prior for user
11. $\alpha^t$is the Dirichlet prior for users at time t.
12. $\beta^t$is the Dirichlet prior for hidden topics at time t.
13. γ is the kernel parameter in the exponential decay function.
14. δ is the size of time interval.
15. w_i^t is the unique word associated with the i-th token of useru at time t.
16. z_i^t is the topic associated with w_i^t.
17. θ is the multinomial distribution of particular topic.
18. θ_u(t) is the multinomial distribution topic specific to the user u at time t.
19. θ_z(t) is the multinomial distribution words specific to the topic z at time t.
20. S be the similarity matrix.
21. St is the users topical similarity matrix at time t.
22. I is the number of iterations in LDA model.
23. M is the keyword matrix.
24. Mt be the users keyword matrix at time t.

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 denotation does not have any effect on users' interests; therefore, we remove it from messages, but keep 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. POS and Keyword Extraction:

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

C. Time Interval Partition:

Users' interests change as time goes by, which reveals and users' microblogs may focus on different topics at different periods of time. Therefore, users' dynamically changing interests can be expressed as a sequence of keyword collections in microblogs at different 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 contains 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 discovering users' interests. As the existence of synonymy, it needs to find the hidden topics from the keyword usage patterns. Since the goal is to find topics that each microblogging user is interested in rather than topics that each microblog is about, we treat the microblogs published by an individual user at the tth time interval as a big document. Then, each row of sub-collection Mt is treated as a bag-of-words document which essentially corresponds to a user. To find user temporal topics in Mt, or to find temporal topics of each 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) associated 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 an additive linear 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 micro blogger 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 in the near future. Generally speaking, users' historical favorites may influence his future interests, andmore recent interests may have 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.

7.CONCLUSION

In this project, we propose a temporal-topic model for friend recommendations in Chinese micro blogging systems. The model first discovers users' latent preferences during different time intervals based on keywords extracted from the aggregated micro blogs 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, followers), the sentiment of micro blogs, 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.

8.REFERENCES

[1] 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.

[2] 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.

[3] 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.

[4] 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.

[5] L. Hong and B. D. Davison, "Empirical study of topic modeling in Twitter," in Proc. 1st Workshop Soc. Media Anal., Washington, DC, USA, 2010, pp. 80–88.

[6] 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.