Scientific Journal of Impact Factor (SJIF): 4.72

e-ISSN (O): 2348-4470
p-ISSN (P): 2348-6406

International Journal of Advance Engineering and Research Development

"Emerging Technologies in the Computer World", January -2017

Friendbook Recommendation

Sanket Ranawade , Rutuja Khedekar , Tejasvi Khedekar , Ajinkya Urankar

(Department of Computer Engineering, AISSMS's Institute of Information Technology)

Abstract--- *Because of the temporary frame and developing ubiquitousness the Microblogging is popping into people's most fascinating alternative for seeking the knowledge and expressing opinions. Messages got by a user in the main admit whom user follows. Therefore, recommending user with comparable interest might enhance the expertise quality for info receiving. Since messages announce by Microblogging users replicate their hobbies or interest and therefore the essential keywords within the messages show their main focus to a large extent, we are able to realize users' preferences by investigation the user generated contents. Besides, user's hobbies, interest aren't static; despite what may well be expected, they modify as time passes by. In lightweight of such instincts, we have a tendency to plan a temporal-topic to investigate user's potential behaviors' and predict their potential friends in Microblogging. The model takes in users' latent preferences by extracting keywords on aggregative messages over a stretch of your time by means that of a subject model, and then the impact of your time is taken into account to deal interest.*

Keywords---*Microblogging, Latent Dirichlet allocation (LDA, CF-Based, Content-Based Hybrid Recommendation Systems.*

I. INTRODUCTION

Microblogging has become a convenient method for web surfers and average users to speak with their friends and members of the family, or to specific intimate emotions or feelings. Employing a microblog additionally has step by step become a habit for an enormous quantity of users that ends up in AN exponential explosion of data within the virtual microblog society on the web, creating retrieving and distinctive required microblog or connected data extraordinarily troublesome. Therefore, a lot of and a lot of microblog services square measure developing novel engines dedicated to recommending user-specific data. Early researchers in the main targeted on the characteristics of Microblogging and social network analysis. Recently, there has been AN increasing interest within the field of data retrieval, like event detection and following, identification of influential individuals, sentiment analysis, and customized recommendations.

Traditional recommendation systems will primarily be classified into 3 categories: CF-based, content-based, and hybrid recommendation systems Probabilistic topic models are well-tried to be the powerful tools for distinctive latent text patterns within the content. Latent Dirichlet allocation (LDA) achieves the capability of generalizing the subject distributions in order that the model may be accustomed generate unseen documents additionally. LDA has additionally been applied to numerous works on Twitter to demonstrate its utility. Users' interests don't seem to be static; contrarily, their interests might amendment as time goes by. Since the time period and brevity options of Microblogging result in frequent updates of microblog, users' interests square measure a lot of intensive and changeable over time.

II. LITERATURE SURVEY

2.1 Paper Name: Topic models towards high performance data mining and analysis
Authors: Katayoun Farrahi, Alois Ferscha
Description: While unimaginable amounts of data are continuously stored recording our transactions, conversations, connections, movements, behaviour, personality, emotions, and opinions, data has been termed "the new oil". The process of"refinement" and knowledge extraction from data is the core of data mining. Advances in automated algorithms and models for extracting knowledge about human behaviour will ultimately measure the value of data. This work discusses the use of probabilistic latent topic models, particularly Latent Dirichlet Allocation (LDA) , for data mining and explores its application on various sorts of large-scale data, focusing on the advantages and disadvantages of their use. While topic models have been shown to provide a promising new tool for data mining, one current open issue is with respect to developing methods for implementing them in high performance computing platforms.

2.2 Paper Name: Friendbook: A Semantic-based Friend Recommendation System for Social Networks
Authors: Zhibo Wang, Jilong Liao, Qing Cao, Hairong Qi,Senior and Zhi Wang
Description: 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. In this paper, we present Friendbook, a novel semantic-based friend recommendation system for social networks, which recommends friends to users based on their life styles instead of social graphs. By taking advantage of sensor-rich smartphones, Friendbook discovers life styles of users from user-centric sensor data, measures the similarity of life styles between users, and

recommends friends to users if their life styles have high similarity. Inspired by text mining, we model a user's daily life as life documents, from which his/her life styles are extracted by using the Latent Dirichlet Allocation algorithm. We further propose a similarity metric to measure the similarity of life styles between users, and calculate users' impact in terms of life styles with a friend-matching graph. Upon receiving a request, Friendbook returns a list of people with highest recommendation scores to the query user. Finally, Friendbook integrates a feedback mechanism to further improve the recommendation accuracy. We have implemented Friendbook on the Android-based smartphones, and evaluated its performance on both small-scale experiments and large-scale simulations. The results show that the recommendations accurately reflect the preferences of users in choosing friends.

2.3 Paper Name: Understanding Transportation Modes Based on GPS Data for Web Applications
Authors: YU ZHENG, YUKUN CHEN, QUANNAN LI, XING XIE and WEI-YING MA
Description: User mobility has given rise to a variety of Web applications, in which the global positioning system (GPS) plays many important roles in bridging between these applications and end users. As a kind of human behaviour, transportation modes, such as walking and driving, can provide pervasive computing systems with more contextual information and enrich a user's mobility with informative knowledge. In this article, we report on an approach based on supervised learning to automatically infer users' transportation modes, including driving, walking, taking a bus and riding a bike, from raw GPS logs. Our approach consists of three parts: a change point-based segmentation method, an inference model and a graph-based post-processing algorithm. First, we propose a change point-based segmentation method to partition each GPS trajectory into separate segments of different transportation modes. Second, from each segment, we identify a set of sophisticated features, which are not affected by differing traffic conditions (e.g., a person's direction when in a car is constrained more by the road than any change in traffic conditions). Later, these features are fed to a generative inference model to classify the segments of different modes. Third, we conduct graph-based post processing to further improve the inference performance. This post processing algorithm considers both the commonsense constraints of the real world and typical user behaviours based on locations in a probabilistic manner. The advantages of our method over the related works include three aspects.

2.4 Friendbook: A Scalable and Efficient Friend Recommendation Using Integrated Feedback Approach
Authors: T. Gayathri Devi and R. Lakshmi
Description: Friend book is a novel semantic-based friend recommendation system for social networks, based on their life styles instead of social graphs which recommend friends to users. Friend book discovers life styles of users, measures the similarity of life styles between users, if their life styles have high similarity it recommends friends to users. User's daily life is modelled as life documents, from which users life styles are extracted by using the Latent Dirichlet Allocation algorithm; Similarity metric to measure the similarity of life styles between users, user's impact is calculated in terms of life styles with a friend-matching graph. A linear feedback mechanism is integrated that exploits the user's feedback to improve recommendation accuracy.

2.4 Paper Name: Introduction to Probabilistic Topic Models
Author: David M. Blei
Description: Probabilistic topic models are a suite of algorithms whose aim is to discover the hidden thematic structure in large archives of documents. In this article, we review the main ideas of this field, survey the current state-of-the-art, and describe some promising future directions. We first describe latent Dirichlet allocation (LDA), which is the simplest kind of topic model. We discuss its connections to probabilistic modelling, and describe two kinds of algorithms for topic discovery. We then survey the growing body of research that extends and applies topic models in interesting ways. These extensions have been developed by relaxing some of the statistical assumptions of LDA, incorporating meta-data into the analysis of the documents, and using similar kinds of models on a diversity of data types such as social networks, images and genetics. Finally, we give our thoughts as to some of the important unexplored directions for topic modelling. These include rigorous methods for checking models built for data exploration, new approaches to visualizing text and other high dimensional data, and moving beyond traditional information engineering applications towards using topic models for more scientific ends

III. PROPOSED SYSTEM

We propose a temporal-topic model to predict user's potential friends. The model 1st extracts user's topic distributions from keyword usage patterns of aggregative messages exploitation temporal approach. Then, it calculates user similarities over time supported user's topic distributions. Finally, users potential interests on others square measure foreseen in keeping 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 users interests; 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.

3.1 Advantageous of planned system

1. Its effective recommendation system for recommending friends to user.
2. It takes less time attributable to the effectiveness of LDA rule.

IV. SYSTEM ARCHITECTURE

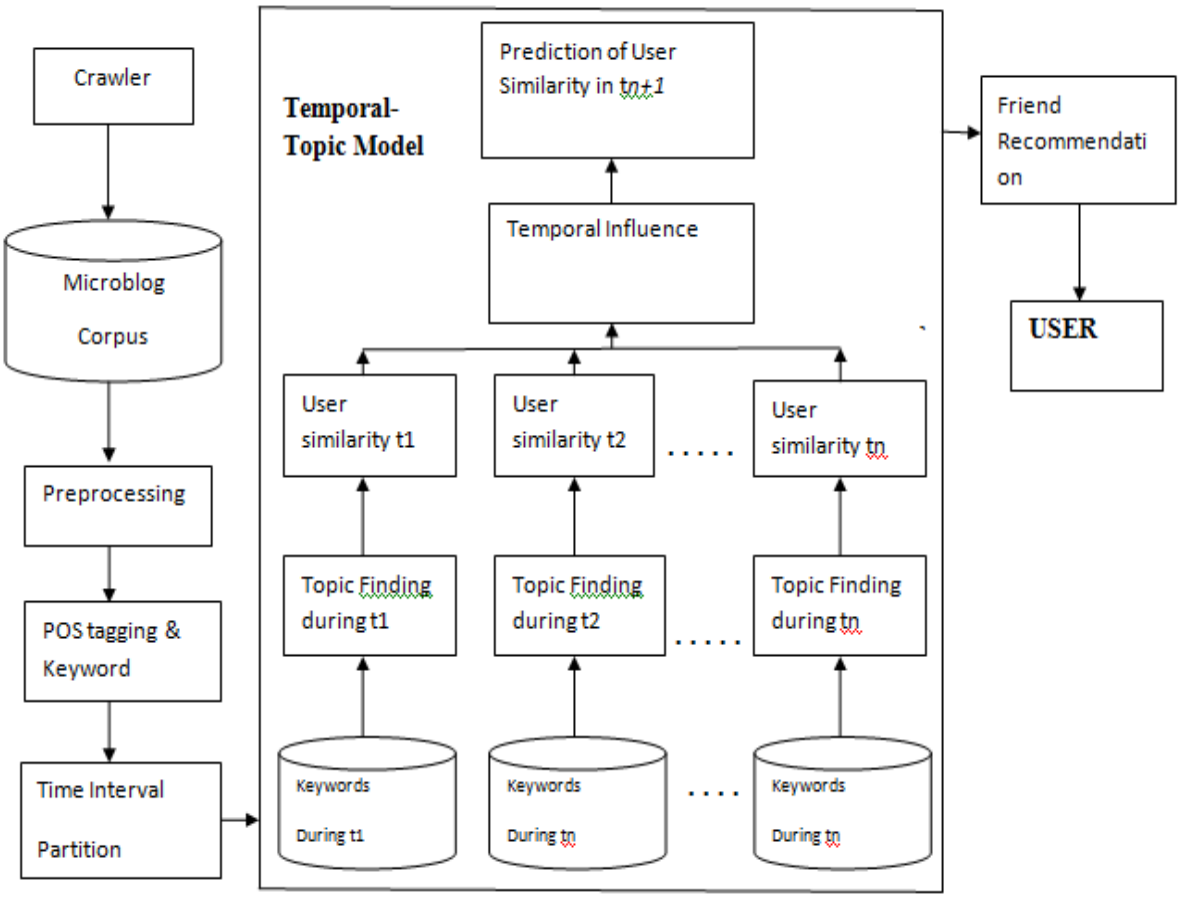

Figure 1. Architecture Diagram of Proposed System

V. 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 intervals.

9. β is the Dirichlet prior for user.
10. α is the Dirichlet prior for user
11. α^t is the Dirichlet prior for users at time t.
12. β^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 users 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.

5.1. 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.

5.2. 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.

5.3. 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 ∪ M2, ... ,∪ 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.

5.4 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.

5.5 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 the 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).

5.6 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; and more 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.

V. CONCLUSION

In this project, we tend to propose a temporal-topic model for friend recommendations in Chinese microblogging systems. The model 1st discovers users' latent preferences throughout totally different time intervals supported keywords extracted from the collective microblogs through a subject model. Then, it calculates user similarities in when interval supported temporal topic distributions. After that, AN decline operate is employed to live interest drifts. Finally, users' potential interests on others are foreseen supported the sequence of users' interests on the timeline. Supported the model, we tend to conducted friend recommendations and therefore the experimental results showed that our model is effective.

For future work, we tend to arrange to conduct our experiments on users WHO have less friends and followers to point out if our model is helpful for the cold-start downside of customized recommendations. We tend to conjointly aim to unearth different factors to boost the performance of the planned model, like social relationships among users (i.e., followers, follows), the sentiment of microblogs, users' location data, etc. we tend to conjointly arrange to investigate different progressive models with temporal evolvement and compare the performances of various ways on friend recommendations. Different datasets like Twitter are tested for the utility and effectiveness of the model.

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.

[7] 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.

[8] 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.

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

[10] 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.

[11] S. Song, Q. Li, and X. Zheng, "Detecting popular topics in microblogging based on a user interest-based model," in *Proc. Int. Joint Conf. IEEE Neural Netw. (IJCNN)*, Brisbane, QLD, Australia, 2012, pp. 1–8.

[12] S. Song, Q. Li, and H. Bao, "Detecting dynamic association among Twitter topics," in *Proc. 21st Int. Conf. Companion World Wide Web*, Lyon, France, 2012, pp. 605–606.

[13] J. Weng, E.-P. Lim, J. Jiang, and Q. He, "Twitterrank: Finding topic-sensitive influential twitterers," in *Proc. 3rd ACM Int. Conf. Web Search Data Min.*, New York, NY, USA, 2010, pp. 261–270.

[14] M. Cha, H. Haddadi, F. Benevenuto, and K. P. Gummadi, "Measuring user influence in Twitter: The million follower fallacy," in *Proc. 4th Int. AAAI Conf. Weblogs Soc. Media (ICWSM)*, Menlo Park, CA, USA, 2010, pp. 10–17.

[15] 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.

[16] 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, SA, 2011, pp. 30–38.

[17] 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.

[18] D. H. Yang and G. Yu, "A method of feature selection and sentiment similarity for Chinese micro-blogs," *J. Inf. Sci.*, vol. 39, no. 4, pp. 429–441, 2013.

[19] H. Bao, Q. Li, S. S. Liao, S. Song, and H. Gao, "A new temporal and social PMF-based method to predict users interests in micro-blogging," *Decis. Support Syst.*, vol. 55, no. 3, pp. 698–709, 2013.

[20] K. Chen *et al.*, "Collaborative personalized tweet recommendation," in *Proc. 35th Int. ACM SIGIR Conf. Res Dev. Inf. Retrieval*, Portland, OR, USA, 2012, pp. 661–670.

[21] J. Chen, R. Nairn, L. Nelson, M. Bernstein, and E. Chi, "Short and tweet: Experiments on recommending content from information streams," in *Proc. 28th Int. Conf. Human Factors Comput. Syst.*, Atlanta, GA, USA, 2010, pp. 1185–1194.

[22] Z. Liu, X. Chen, and M. Sun, "Mining the interests of Chinese microbloggers via keyword extraction," *Front. Comput. Sci.*, vol. 6, no. 1, pp. 76–87, 2012.

[23] F. Abel, Q. Gao, G.-J. Houben, and K. Tao, "Semantic enrichment of Twitter posts for user profile construction on the social web," in *Proc. 8th Extended Semantic Web Conf. Semantic Web Res. Appl.*, Berlin, Germany, 2011, pp. 375–389.

[24] Z. Xu, Y. Zhang, Y. Wu, and Q. Yang, "Modeling user posting behavior on social media," in *Proc. 35th Int. ACM SIGIR Conf. Res. Dev. Inf. Retrieval*, Portland, OR, USA, 2012, pp. 545–554.

