Semantic Topic-Modelling Methodologies for Friend Recommendation based on lifestyles

Kandiraju Sai Ashritha
Department of Computer Science
and Engineering
National Institute of Technology
Karnataka, Surathkal
Email: 14co121@nitk.edu.in

***Abstract*—Nowadays, Social Networks are appreciating a noteworthy increment in fame. Online social networking not only helps to keep in touch with friends and family, it also helps in building up virtual groups and establishing online relationships. The virtual-friendship over networking sites can be substantially influenced by the nature and quality of recommendations. Most of the social networking sites rely on pre-existing relationships and adapt a friends-of-friends approach to recommend potential candidates. Individuals who share common interests, tastes, habits or lifestyles tend to connect at a faster pace virtually. Thus taking into account the lifestyles of users while recommending friends can be more meaningful and intuitive. Discovering user lifestyles from their life documents can be achieved through traditional probabilistic topic models like Latent Dirichlet Algorithm. But lifestyle extraction using LDA algorithm is not entierly realistic as LDA doesn't take into account important things like sentence structure, correlation between topics, short life documents into account. This paper proposes methodologies to obtain user lifestyles from life documents in the semantic sense.**

Keywords*—*Friend recommendation, lifestyles, life documents, topic-modelling, LDA, TextRank, CTM, BTM

I. Introduction

Nowadays, Social Networks are appreciating a noteworthy increment in fame. The primary reason for this popularity is that people stay connected through the cyber world. In today's widely connected world, majority of the population use social networking as an essential means of communication with their loved ones. Services such as Facebook, Twitter, LinkedIn and Google+ are amongst the most popular online social networks.

Statistics infer that users of the social networking platforms tend to connect with people well-acquainted in real world and also with individuals who are not by any means familiar. Online social networks have revolutionized the way people make friends. Facebook provides its users with innumerable opportunities to "Friend" other users of the networking platform [1]. This could change the way in which individuals approach friendship. Friendship beforehand required time, exertion, and a considerable measure of discussion, it now requires just a click of a button i.e users "friend" others without much of a speculation.

Online social networking not only helps to keep in touch with friends and family, it also helps in building up virtual groups and establishing online relationships [2]. The virtual-friendship over networking sites can be substantially influenced by the nature and quality of recommendations. Therefore, friend recommendation and good suggestions for friends is a major issue in online social networks. A large portion of the recommendation techniques depend on previous user connections to pick candidates for suggestion. For instance, Facebook depends on a social link connections among the individuals share common friends and suggests symmetrical users as friends. In a similar manner most of the networking sites rely on pre-existing relationships and adapt a friends-of-friends approach to recommend potential candidates. But recent research has shown that this approach is not very suitable to today's modern day virtual friendships [3].

Individuals who share common interests, tastes, habits or lifestyles tend to connect at a faster pace virtually. Thus taking into account the lifestyles of users while recommending friends can be more meaningful and intuitive. Thus, if we could obtain data about user lifestyles through their day-to-day activities and daily routines, we can effectively recommend friends based on their lifestyles.

The paper is oraganised as follows: Section II provides the related work on recommendation systems. Section III elaborates the problem being tackled and the need for addressing this problem. Section IV provides methodologies in efficiently extracting the lifestyles of users from available user data. Section V provides a brief summary and concludes the paper.

II. Literature Review

Recommendation systems have attracted a huge amount of attention from the research community. It has always been an active area of research with significant proposals and contributions. In [4], the authors have presented a recommendation system that takes physical and social context into consideration. In [5], the authors have proposed MatchMaker, a filtering friend recommendation system that

matches people based on their personality. The authors of [6], have put forward a recommendation system that suggests friends based on their geographical locality information obtained using the GPS technology. In [7], W. H. Hsu et al. have proposed a collaborative recommendation approach that exploits the link structure of the social networks. Also they have used content-based recommendation in the paper by taking the mutual interests of the users into consideration. However, the above mentioned recommendation systems do not give any importance to user's day-to-day activities.

Various types of wearable sensors have been used to obtain low-level sensor data in an attempt to understand the user's daily routines and activities. The authors in [8], have made use of Hidden Markov models to understand user activities from the data obtained from the wearables. In the paper [9], the authors have made use of changing GPS data to understand the mode of commute of different users. Li et al. [10] made use of accelerometers and gyroscopes inorder to recognize static postures and dynamic transitions of the users.

Not just wearbles, smartphones these days are an invaluable resource in obtaining sensor rich data. The authors in [11], have used the inbuilt accelerometer and GPS technologies of a smart phone to identify the mode of transportation of the user. Soundsense [12], has obtained information about user specific sounds(eg: music,voice etc.) with the help of the microphone on the smartphone. In this way significant amount of work has been done in order to extract day-to-day activities of the users with the help of sensor data from a smartphone. Relatively less research has been done in extracting user lifestyles or daily routines from smartphones. MIT Reality Mining Project [13] made an attempt to obtain location-driven routines from GPS data. Routines like leaving for work and stopping by a restaurant on the way to have a meal could be successfully inferred. A drawback worth pointing out would be that the system failed in capturing the routines of users who remain static i.e users whose location doesnt change.

In Friendbook [14], the authors have presented a novel-semantic based recommendation system that suggests friends based on user lifestyles. The user data is obtained from the sensor-rich smartphones and the user's lives are modelled as *life documents*. It is stated that a user's day-to-day life is comprised of several activities like "walking","eating" etc. These activities together form daily routines like office work, shopping, eating at a cafe etc. For example, eating at a cafe would include activities like walking to the cafe, sitting and eating. This way the authors drew an analogy between people's lives and the life documents that have been obtained from the smartphones. The life documents are assumed to be made up of lifestyles(daily routines) and lifestyles are assumed to be made up of activies as shown in Fig. 1. The authors have extracted user lifestyles from their life documents using the Latent Dirichlet Allocation (LDA) algorithm. The lifestyles thus obtained have been compared and the users with maximum similarity in their lifestyles have been recommended to each other. Another paper [15],

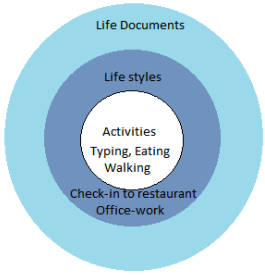

Fig. 1. Analogy between Life Documents and User daily lives

works like a client-server application where life documents of users have been collected through the browser. The user lifestyles were extracted using LDA algorithm, SQL and hadoop technologies. In this way the system recommends users with similarity in location, blood group, interest etc.

III. Problem Statement

As already mentioned Friendbook [14] and [15] have adopted probabilistic topic model to discover the probabilities of the lifestyles from the life documents of the users. In probabilistic topic models the frequency of words is important. A bag-of-activity model can be then proposed by replacing the original sequence of activities with their probability distributions as shown in Fig. 2. This way each user's life document is transformed into a bag-of-activity representation comprising of his day-to-day activities. Let

$$w = [w_1, w_2, ..., w_W]$$

denote the set of activities where w_i is the i^{th} activity and N is the total number of activities. Let

$$z = [z_1, z_2, ..., z_N]$$

denote the set of lifestyles where z_i is the i^{th} lifestyle and Z is the total number of lifestyles. Let

$$d = [d_1, d_2, ..., d_N]$$

denote the set of life documents where d_i is the i^{th} life document and N is the total number of users. Let $p(w_i \mid d_k)$ denote the probability with which a certain activity w_i occurs in life document d_k. Let $p(w_i \mid z_j)$ denote the probability of how much a particular activity w_i contributes to the lifestyle z_j. Let $p(z_j \mid d_k)$ denote the probability of a lifestyle z_j being a part of the life document d_k.

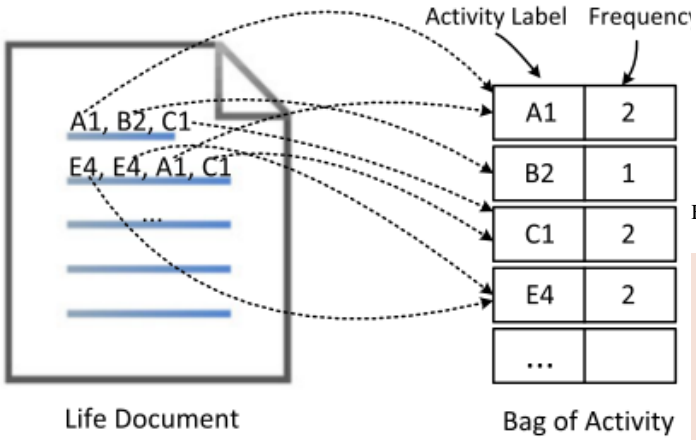

Fig. 2. Bag-of-Activity Model

Each user's lifestyle is denoted by a lifestyle vector denoted by, L_k = [$p(z_1 \mid d_k, p(z_2 \mid d_k, ..., p(z_Z \mid d_k$]. The authors aimed to obtain these lifestyle vectors using the Latent Dirichlet Allocation Algorithm.

A. Life Style Extraction using LDA Algorithm

The LDA model can be represented using the equation

$$p(w|d) = p(w|z)p(z|d)$$

where $p(w \mid d) = [p(w \mid d_1, p(w \mid d_2, ..., p(w \mid d_N]$ is the activity-document matrix as shown in Fig. 3. which constitutes the probability of each activity over each life document and $p(w \mid d_k) = [p(w_1 \mid d_k, p(w_2 \mid d_k, ..., p(w_W \mid d_k]^T$ is the k^{th} column in the activity-document matrix denoting the probabilities of all the activities over a particular life document d_k of a k^{th} user. $p(w \mid z) = [p(w \mid z_1, p(w \mid z_2, ..., p(w \mid z_Z]$ represents the activity-topic matrix as shown in Fig. 3. depicting the probability of each activity over each lifestyle and $p(w \mid z_k) = [p(w_1 \mid z_k, p(w_2 \mid z_k, ..., p(w_W \mid z_Z]^T$ is the k^{th} column in the activity-topic matrix denoting the probabilities of all the activities over a particular lifestyle(topic) z_k. $p(z \mid d) = [p(z \mid d_1, p(z \mid d_2, ..., p(z \mid d_N]$ represents the topic-document matrix as shown in Fig. 3. It contains the probability of each lifestyle over each life document and $p(z \mid d_k) = [p(z_1 \mid d_k, p(z_2 \mid d_k, ..., p(z_Z \mid d_k]^T$ is the k^{th} column in the topic-document matrix denoting the probabilities of all lifestyles over a particular life document d_k. The algorithm works as mentioned in [16] and Fig. 4. provides an intuition of LDA assigns probabilities.

In this way, LDA algorithm has been used for extraction of the lifestyles of users. But LDA has its own set of disadvantages which affect the extracted life style feature vectors which in turn affect the recommendation system. The problems of LDA algorithm worth mentioning in this context

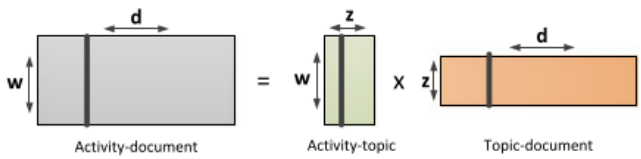

Fig. 3. LDA model

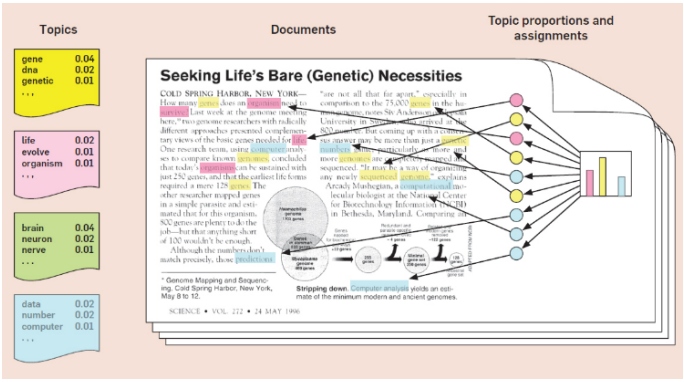

Fig. 4. Intuition behind working of LDA algorithm. On the left, assumed topics from the document are shown. On the right, the probabilities of the topics in the document are shown.

would be:

- The number of topic that need to be modelled should be known prior to the execution of the algorithm. K is fixed.
- Latent Dirichlet Allocation algorithm cannot capture correlations between the topics.
- The algorithm assumes a Bag-of-Words approach. The model assumes that words are exchangeable and therefore capturing the sentence structure is not possible.
- LDA algorithm performs poorly in environments containg short-texts.

The above mentioned problems would affect the way in which the lifestyles of the users are captured. This paper therefore suggests other semantic topic-modelling methodologies that could be employed in obtaining lifestyles(topics) from user life documents.

IV. SEMANTIC TOPIC-MODELLING METHODOLOGIES

A. Keyword Extraction using TextRank

Consider two students S1 and S2. Let their life documents be L1 and L2.
L1 = "Student named XYZ. Studies in PQR university. Has taken the course Integral calculus. Feels integral calculus is the most powerful branch of mathematics. Impressed with

how calculus works with varying quantities. Uses calculus to find areas under curve etc. ”
L2 = “Student in PQR university with name XYZ. Favorite subject is differential calculus. Uses differential calculus which is a sub-field of calculus to calculate rate of change of quantities.”

From the above life documents we understand that the two students though have calculus in common, actually love two different branches of calculus. Students who prefer differential calculus should be recommended with more preference to S2 than recommending S1 to S2. Bag-of-Activity model when constructed for both the students has the word **calculus** with maximum frequency. When LDA is applied over these models, the lifestyle vectors obtained would have “calculus” with the maximum probability. This increases similarity between both the lifestyles and the recommendation system recommends S1 to S2. The reason for this is, LDA assumes a bag-of-activity approach. That is, it does not consider the relationship between the words. This can be overcome if important phrases from the documents are captured instead of just words(activities). For example, brisk walking, classical singing, ballet dancing, eating in XYZ restaurant should be given more weightage than walking, singing, dancing, eating respectively. The latter are more generic whereas the latter are more user specific and can be used in capturing the lifestyles. TextRank can be used to extract phrases from a life documents using graphical approach. Probabilistic topic models can be then used to obtain the probabilities for the construction of lifestyle vectors.

TextRank [17] is an unsupervised keyword extraction algorithm which uses graphs for ranking of words. In any graph-based ranking model, the idea is to take “votes” from vertices i.e when one vertex is joined to another vertex through an edge, this means that it is casting a vote for the other vertex. A vertex gains importance with the increasing number of votes that have been cast for it. Also, the importance of a vertex casting a vote determines the importance of the vote, therefore a vertex having many important votes will be of maximum importance. Thus, a score calculated for each vertex depends on the number of votes that have been cast for it as well as the scores of the vertices casting those votes. At the start of the algorithm, random values are assigned to the nodes of the graph and the algorithm iterates till convergence is achieved. After running the algorithm, the scores for each vertex are obtained. These scores explicitly depict the “importance” of a given vertex in the graph. A point to be noted is that the final scores are not affected by the choice of the initial arbitrary values. The only difference would be in the number of iterations that need to be performed in order to converge.

When applying TextRank to process natural language texts, we need to construct a graph that represents the text and the meaningful relations in the text. For keyword extraction using TextRank we build a graph by taking the vertices of the graph as words. If two words are related, then we place an edge between the vertices representing those words in the graph.

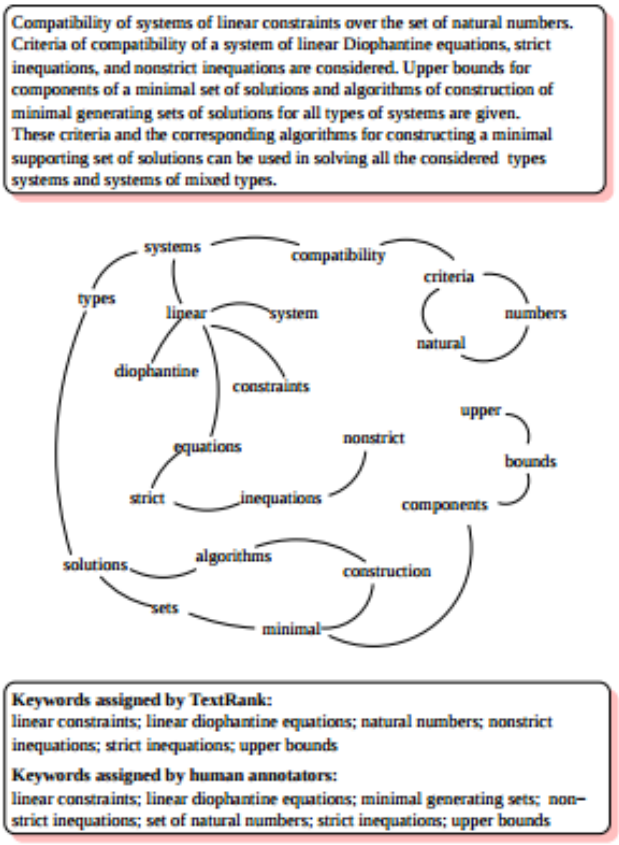

Fig. 5. Keywords extraction using TextRank

The relation used here is a **co-occurrence relation**, which depends on the distance between the word occurrences. Two words are said to be connected only if their words co-occur within a window of maximum N words and the N value can vary from 2 to 10. Semantic relations between the words can thus be captured using the co-occurrence links.

The algorithm proceeds as follows : Token are obtained from the text and their parts of speech are identified. This preproccesing step ensures that the algorithm works with words of specified parts of speech. Single words from the text are added to the graph and multi-word keywords will be obtained after running the algorithm and postprocessing the output. An edge is added between all the words that belong to the window of words (whose size is pre-determined). The graph is constructed with initial values of all the vertices set to 1. The algorithm is run until it converges (usually 20-30 iterations). The final scores of the vertices are sorted in the reverse order and the vertices on the top are retained. These vertices if present adjacent to each other, are collapsed to form a multi-word keyword and these words are returned as the potential keywords that have been extracted. Fig 5. shows the graph built and keywords extracted for the given abstract.

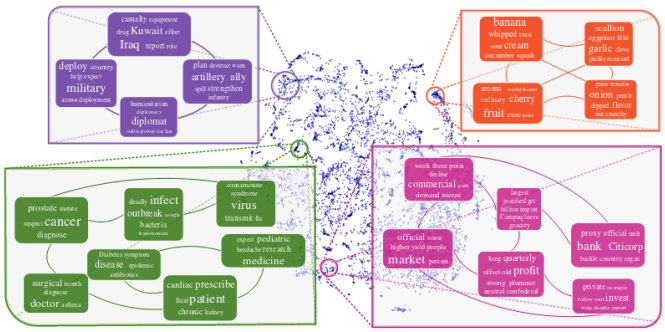

Fig. 6. Topic-modelling using Correlated Topic Models

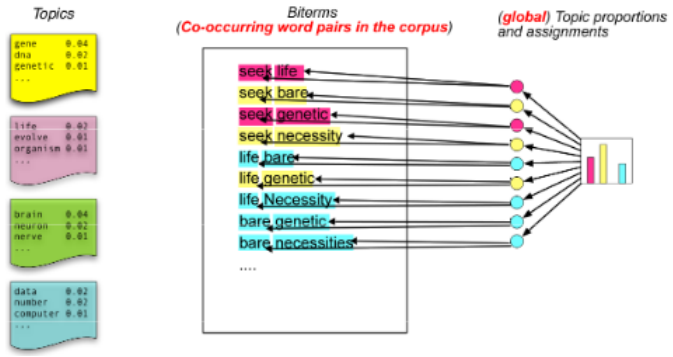

Fig. 7. Topic-modelling using Correlated Topic Models

B. Correlated Topic Models (CTM)

LDA algorithm models topic occurrence almost independently. As a result it fails to capture rich correlations among the topics. For example, a user lifestyle about drama is more likely to be correlated to entertainment than to finance. Thus effective modelling of the topics by taking their structural correlations into account is extremely essential for improved lifestyle extraction. The strong independence between topics assumed in LDA is not practical and realistic. Correlated Topic Model (CTM) proposed in [18] is an extension of LDA which replaced the Dirichlet with the logistic-normal prior. This was effective in capturing the pairwise topic correlations with the Gaussian co-variance matrix. Talking technically, a topic model is a generative probabilistic model that makes use of a few distributions over a set of words to describe a document collection. When fit from data, the distributions would actually correspond to intuitive notions of topicality. Fig. 6. shows the topics modelled from NYTimes news corpus using CTM and the correlations between the topics.

C. A Biterm Topic Model for Short Texts

Not just sensor data, user's lifestyles can also be captured from their web actions. The bio of an individual in Twitter, Instagram, Facebook and other social networking platforms can be used to understand his/her lifestyles. Short texts like search queries, tweets, bio and comments however suffer from data sparsity and so traditional topic-modelling algorithms like LDA fail in this context. Short text snippets do not contain enough number of words for the model to learn the relations between the words. Unline proper documents, it is also difficult with short texts to disambiguate multiple meanings of a single word. Problems faced when LDA is used with short texts :

- Word counts are not discriminative - In a normal document, words belonging to a particular topic occur frequently. On the other hand, in short texts most words only occur once.
- Not enough contexts to identify the sense of ambigous words - In a normal document with rich context, we have many relevant words to have an idea about the context. Whereas in a short text, achieving this becomes difficult as the context in limited with only few relevant words.

Second, it supposes each biterm is draw from a topic. Inferring the topic of a biterm is also easier than inferring the topic of a single word in LDA, since more context is added.

The key idea of Biterm Topic Model (BTM) [19] is to model the whole corpus as a mixture of models. The reason is inferring topic mixture over an entire corpus is easier than inferring the topic mixture over a short text. Next it supposes that each biterm is a draw from a topic. As mentioned above, it is easy to infer the topic of a biterm than that of a single word in LDA. The reason being increase in the amount of context added. Fig. 7. shows the result after applying Biterm Topic Model.

V. Conclusion

The paper presented the need for recommendation systems based on user lifestyles. The algorithms used for extracting life documents have been explained elaborately. To overcome the disadvantages faced due to lifestyle extraction from life documents using LDA, other topic-modelling approaches have been proposed. The proposed methodologies take into account the semantic nature of the life documents before obtaining lifestyle feature vectors.Topic-models like TextRank, Correlated Topic Model (CTM) and Biterm Topic Model (BTM) have been explained with respect to semantic lifestyle extraction.

References

[1] Facebook statistics. (2011). [Online]. Available: http://www.digitalbuzzblog.com/facebook-statistics-stats-facts-2011/

[2] Xing Xie., Potential Friend Recommendation in Online Social Network, 2010 IEEE/ACM International Conference on Green Computing and Communications 2010 IEEE/ACM International Conference on Cyber, Physical and Social Computing,2010.

[3] M. Tomlinson, Lifestyle and social class, Eur. Sociol. Rev., vol. 19, no. 1, pp. 97111, 2003.

[4] J. Kwon and S. Kim, Friend recommendation method using physical and social context, Int. J. Comput. Sci. Netw. Security, vol. 10, no. 11, pp. 116120, 2010.

[5] L. Bian and H. Holtzman, Online friend recommendation through personality matching and collaborative filtering, in Proc. 5th Int. Conf. Mobile Ubiquitous Comput., Syst., Services Technol., 2011, pp. 230235.

[6] X. Yu, A. Pan, L.-A. Tang, Z. Li, and J. Han, Geo-friends recommendation in GPS-based cyber-physical social network, in Proc. Int. Conf. Adv. Social Netw. Anal. Mining, 2011, pp. 361368.

[7] W. H. Hsu, A. King, M. Paradesi, T. Pydimarri, and T. Weninger, Collaborative and structural recommendation of friends using weblog-based social network analysis, in Proc. AAAI Spring Symp. Ser., 2006, pp. 5560.

[8] J. Lester, T. Choudhury, N. Kern, G. Borriello, and B. Hannaford, A hybrid discriminative/generative approach for modeling human activities, in Proc. Int. Joint Conf. Artif. Intell., 2005, p. 766- 772.

[9] Y. Zheng, Y. Chen, Q. Li, X. Xie, and W.-Y. Ma, Understanding transportation modes based on GPS data for web applications, ACM Trans. Web, vol. 4, no. 1, pp. 136, 2010.

[10] Q. Li, J. A. Stankovic, M. A. Hanson, A. T. Barth, J. Lach, and G. Zhou, Accurate, fast fall detection using gyroscopes and acceler- ometer-derived posture information, in Proc. 6th Int. Workshop Wearable Implantable Body Sensor Netw., 2009, pp. 138143.

[11] S. Reddy, M. Mun, J. Burke, D. Estrin, M. Hansen, and M. Srivastava, Using mobile phones to determine transportation modes, ACM Trans. Sens. Netw., vol. 6, no. 2, p. 13, 2010.

[12] E. Miluzzo, C. T. Cornelius, A. Ramaswamy, T. Choudhury, Z. Liu, and A. T. Campbell, Darwin phones: The evolution of sens- ing and inference on mobile phones, in Proc. 8th Int. Conf. Mobile Syst., Appl., Services, 2010, pp. 520.

[13] N. Eagle and A. S. Pentland, Reality mining: Sensing complex cocial systems, Pers. Ubiquitous Comput., vol. 10, no. 4, pp. 255 268, Mar. 2006.

[14] Zhibo Wang, Hairong Qi, Friendbook: A Semantic-Based Friend Recommendation System for Social Networks, IEEE Transactions on Mobile Computing, Vol. 14, No. 3, MARCH 2015.

[15] T. R. Kacchi and A. V. Deorankar, "Friend recommendation system based on lifestyles of users," 2016 2nd International Conference on Advances in Electrical, Electronics, Information, Communication and Bio-Informatics (AEEICB), Chennai, 2016, pp. 682-685

[16] D. M. Blei, A. Y. Ng, and M. I. Jordan, Latent Dirichlet allocation, J. Mach. Learn. Res., vol. 3, pp. 9931022, 2003

[17] R. Mihalcea and P. Tarau. 2004. TextRank bringing order into texts

[18] David M Blei and John D Lafferty.2007, A correlated topic model of science. *The Annals of Applied Statistics* (2007),17-35

[19] Yan, X., Guo, J., Lan, Y., and Cheng, X. (2013). A biterm topic model for short texts. In Proceedings of the 22nd international conference on World Wide Web , WWW 13, pages 14451456, Republic and Canton of Geneva, Switzerland. International World Wide Web Conferences Steering Committee.