A Better Semantic based Friend Recommendation System for Modern Social Networks

Akash Bhapkar
Dept. of Computer Engg
AISSMS COE
Pune, India

Kajal Fegade
Dept. of Computer Engg
AISSMS COE
Pune, India

Rahul Ahire
Dept. of Computer Engg
AISSMS COE
Pune, India

Chitra Chaudhary
Dept. of Computer Engg
AISSMS COE
Pune, India

A. M. Jagtap
Dept. of Computer Engg
AISSMS COE
Pune, India

ABSTRACT

People meet new people on social networks from across the world, eventually bringing the world closer. Existing social systems recommend friends according to person's social graph which includes mutual friends and social connections, which may not be the case when being friends in real life. A better approach would be to get recommendations according to user's life-style rather than just social graph. The proposed system provides an intelligent and automated way to predict user's lifestyle according to his daily activities and interests by taking advantage of sensor-rich smart-phones and recommend friends with high similarity of lifestyles. The user's data is stored in database and lifestyle is extracted using topic model. By constructing friend-matching graph, our system depicts the similarity of lifestyles between two users. Upon receiving a request, a list of people with highest recommendation scores is returned to the query user. Finally, system integrates a feedback mechanism to further improve the recommendation accuracy.

Keywords

Friend Recommendation, text mining, lifestyle, mobile sensing, social networks

1. INTRODUCTION

Social networking sites have made to every corner of the world. Today, with 1700 million users on Facebook, the hunger to connect the social world is perpetual. Tonnes of people meet tonnes of others in this network, which undoubtedly brings the world closer. Although users for social networks has increased, the big reason behind it was the increase number of smart-phone users. Just 8 years back, smart-phones with android system was launched and changed the face of the world till now. Report says that more than 2000 million users access social networking sites from their smart-phones. Moreover, an average person spends 3-4 hours a day in his smart-phone that too without having any awareness of the surrounding. This has led the increase in ease of access of the sites and eventually increase in number of users, as everything needed is fitted only in one hand. Today's smart-phones come with variety of very rich sensors which are used for the human benefits only. In short, smart-phones have revolutionized the world.

In order to detect the accurate lifestyle of a user using his activity, there was no choice but the wearable sensors. But now, with sensor-rich smart-phones it is possible to predict the user's lifestyle with reasonable characteristics. By connecting social network with smart-phone would be preferable, as smart-phone will detect user's lifestyle and social networks will provide friend recommendations accordingly. This will surely result as a better approach.

2. LITERATURE SURVEY

In 2015, Z. Wang, J. Liao, Q. Cang, H. Qi, Z. Wang represented a semantic based approach to recommend friends to to user. By taking advantage of sensor-rich smart-phones, Friendbook discovered life styles of users from user-centric sensor data, measured the similarity of life styles between users, and recommended friends to users if their life styles have high similarity. Inspired by text mining, they modelled a user's daily life as life documents, from which his/her life styles were extracted by using the Latent Dirichlet Allocation algorithm. Further, they proposed a similarity metric to measure the similarity of life styles between users, and calculate user's 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 integrated a feedback mechanism to further improve the recommendation accuracy. They have implemented Friendbook on the Android-based smart-phones, and evaluated its performance on both small-scale experiments and large-scale simulations. The results show that the recommendations accurately detected the preferences of users in choosing friends. This view very closely aligns with our system, but we are going a step further with locations and app-usages in the smart-phones.

In 2011, L. Bian and H. Holtzman, presented "Matchmaker", a collaborative filtering friend recommendation system based on personality matching[2]. The Matchmaker aimed to leverage the social understanding and mutual understanding among the people in the existing social network connections and recommend friends based on rich contextual data from people's physical world interactions. MatchMaker allowed user's network to match them with similar TV characters, and used relationships in the TV programs as a parallel comparison matrix to suggest to users friends that have been voted to suit their personality the best. The system's ranking schema allowed progressive improvement in the personality matching consensus and user's social network connections more diversely branched.

In 2011, J. Kwon and S. Kim, proposed a friend recommendation method based on physical and social context. In the paper [3], the main idea of the proposed method consisted of the following three stages- 1. computing the friendship score using physical context; 2. computing the friendship score using social context; 3. combining all of the friendship scores and recommending friends by the scoring values. The context-aware computing environments were considered. To score the friendship both the spiritual friendship and social friendship were considered. Physical contexts were used for scoring the spiritual friendship and the social contexts for the social friendship. After getting the friendship score, strength between the user and user's friend was computed and the friends were recommended in ascending order.

In 2011, X. Yu, A. Pan, L.-A. Tang, Z. Li, and J. Han, aimed to recommend geographically related friends which could help wed-based social service users to find more friends in the real world. They combined the GPS information and social network structures, to build a pattern based heterogeneous information network. Links inside this network reflected both people's geographical information, and their social relationships. The estimated link relevance found promising geo-friends by employing a random walk process on the heterogeneous information network. In paper [4], friend recommendation problem in cyber-physical social network was studied. With location and trajectory information available, improved the accuracy of the results and make on-line social services much closer to users' real life.

In 2011, K. Farrahi and D. Gatica-Perez, discovered the daily location-driven routines which are contained in a massive real life human dataset collected by mobile phones [5]. They aimed discovery and analysis of human routines which would characterize both individual and group behaviours in terms of location patterns and developed an unsupervised methodology based on two differing probabilistic topic models and applied them to the daily life of 97 mobile phone users over a 16 month period to achieve these goals. Topic models are probabilistic generative models for documents that identify the latent structure that underlies a set of words [5]. Routines dominating the entire group's activities, identified with a methodology based on the Latent Dirichlet Allocation topic model, include going to work late, going home early, working non-stop and having no reception (phone off) at different times over varying time-intervals. They were also able to characterize daily patterns by determining the topic structure of days in addition to determining whether certain routines occur dominantly on weekends or weekdays.

In 2010, K. Farrahi and D. Gatica-Perez. suggested that human interaction data, or human proximity, obtained by mobile phone Bluetooth sensor data, can be integrated with human location data, obtained by mobile cell tower connections, to mine meaningful details about human activities from large and noisy datasets [6]. A model was proposed, called as bag of multi-modal behaviour that integrated the modelling of variations of location over multiple time-scales, and the modelling of interaction types from proximity. They used an unsupervised approach, based on probabilistic topic models, to discover latent human activities in terms of the joint interaction and location behaviours of 97 individuals over the course of approximately a 10-month period using data from MIT's Reality Mining project. By computing the entropy of individuals based on their jointly modelled locations and interactions, their method was able to predict missing multi-modal data over several hours for users with both low and highly varying lifestyles.

In 2008, Huynh, Fritz and Schiel introduced an approach for modelling and discovering daily routines from on-body sensor data. Inspired by machine learning methods from the text processing community, they converted a stream of sensor data into a series of documents consisting of sets of discrete activity labels. These sets are then mined for common topics, i.e. activity patterns, using Latent Dirichlet Allocation. In an evaluation using seven days of real-world activity data, they showed that the discovered activity patterns correspond to high-level behaviour of the user and are highly correlated with daily routines such as commuting, office work or dinner routine. The patterns can be based on a learned vocabulary of meaningful activity labels (such as walking, using the phone, discussing at whiteboard, etc.), in which case the discovered patterns are immediately human-readable in that they represented sets of such labels.

3. PROPOSED SYSTEM

(1) Semantic based Friend recommendation system primarily focuses on recommendations based on lifestyle similarity among the users.

(2) On the client side, smart-phone of each user will record data of its user and report the generated life documents to the server.

(3) For better analysis of lifestyle, we incorporate the data from smart-phones such as application usage and location preferences which eventually reflects user's lifestyle more accurately.

(4) Our system analyses the data stored in life documents collected from user's phone and based on that analysis, a user is indexed as per the lifestyle recognized.

(5) Then Friend Matching Graph will be constructed and top users with similar lifestyle with a particular user will be recommended. Using this friend matching graph user's ranking is decided accordingly.

(6) Based on user's query, the top number of users will be recommended to the user. Further, our system incorporates the feedback control mechanism, check whether the recommended list was accurate and useful to user or not.

3.1 Architecture

In this section we give the brief overview of the system architecture of our system. Figure 2 shows system architecture which adopts a client-server model where each client is a user's smart-phone and the servers are data centres or clouds.

3.1.1 Data Collection and Data Analysis:. The Collection module collects life documents from the user's smart-phone and stores it into a file in either semi-structured or structured format. After collecting the documents, the life styles of users are extracted by the life style analysis module. It may use Hadoop technology or SQL technology depending on the type of file as input to it.

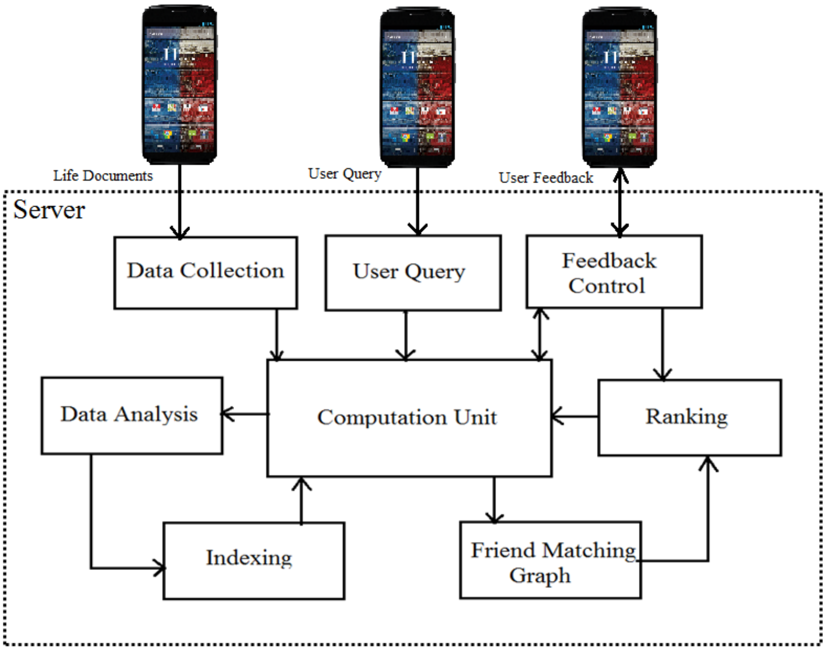

Fig. 1. System Architecture

3.1.2 Indexing:. After analysing the life style of a particular user, indexing module puts the input data of life styles of users into the database in the format of (life-style, user) instead of (user, lifestyle). This eventually decreases the overhead of searching every user's different lifestyle and we get the total number of users living the similar lifestyle.

3.1.3 Friend-Matching Graph:. This graph can be constructed accordingly with the help of graph data structure construction module. With the help of this Friend-Matching graph, we analyse the similarity between users. Nodes are users and edges represents the similarity of lifestyles, edge is present among two nodes if it satisfies the threshold probability of lifestyle.

3.1.4 Ranking:. After graph is constructed, the ranks of users are then calculated. The higher the number of edges per node or user, the higher is the rank. Recommendation quality depends on two things, first is the similarity between the users and second is user's ranking. The top users in the recommendation provided will have higher similarity measure and ranking.

3.1.5 User Query:. This module takes a user?s request and shows a list of recommended friends to the user. Query includes the decision factor which will be useful to determine which users will be recommended as friends.

3.1.6 Feedback Control:. This module takes care of the feedback given by the user according the recommendation list provided by the system. This will be helpful to improve quality of recommendation and to meet user's satisfaction.

3.2 Compute Similarity and Ranking

To compute the similarity among two users was a hurdle task as the beginning. But after representing user's lifestyle in a feature vector, the similarity computation became easier. LDA takes the life documents of users and after calculating the activity frequencies in the document, probabilities of every activities with respect to lifestyle is also calculated. Now, these probabilities will be served as a feature vector of a particular user. We have defined the number of lifestyles as 7 and hence our feature vector contains 7 probabilities.

The similarity of life styles between user i and user j, denoted by Similarity(i,j), is defined as follows:

$$Similarity(i, j) = Similarity_c(i, j)Similarity_d(i, j)$$

where $Similarity_c(i, j)$ is also known as cosine similarity is used to measure the similarity of the life style vectors of users as a whole, $Similarity_d(i, j)$ is also known as dominant lifestyle similarity is used to emphasize the similarity of users on their dominant life styles.

The cosine similarity is calculated as

$$Similarity_c(i, j) = cos(Li, Lj)$$

Now, for dominant lifestyle similarity, users feature vector is analysed and the dominant lifestyles are extracted. Dominant lifestyle are those which satisfies the following requirements:

(1) The total probability distribution of the set is larger than or equal to λ which is a predetermined threshold.

(2) The probability distribution of any life style in the set is larger than or equal to that of any life style not in the set.

(3) The set should have the minimum number of life styles.

Therefore, we get the set $Dominant_i$ in which the dominant lifestyles of user i are stored. The similarity metric $Similarity_d(i, j)$ for measuring the similarity of the dominant life style sets of two users is then defined as

$$Similarity_d(i, j) = \frac{2.|Dominant_i \cap Dominant_j|}{|Dominant_i| + |Dominant_j|}$$

Since both $Similarity_c(i, j)$ and $Similarity_d(i, j)$ vary between 0 and 1, we conclude that the similarity metric $Similarity(i, j)$ varies between 0 and 1

As an example to show the calculation of two user's life style similarity, we assume that there are two users 1 and 2 in the system, who have the life style vectors $L_1 = [0.3, 0.1, 0.3, 0, 0.1, 0.2, 0.1]$ and $L_2 = [0.1, 0.3, 0.2, 0, 0.2, 0, 0.1]$ respectively. The number of life style topics is 7. We first calculate $Similarity_c(1, 2) = cos(L_1, L_2) = 0.688$. Given = 0.8, we can calculate the dominant life style sets of these two users, $D1 = z1, z3, z6$ and $D2 = z2, z3, z5, z1$ respectively. Therefore, the dominant life style similarity is calculated as $Similairty_d(1, 2) = \frac{2x2}{3+4} = 0.571$. Finally, the similarity of user 1 and 2 is $Similarity(1, 2) = Similarity_c(1, 2)Similarity_d(1, 2) = 0.393$.

Using above similarity measures, the friend matching graph is constructed which contains user as nodes and their similarity as edges between them. If the similarity among two nodes or users is greater than or equal to the thresh-hold similarity then that edge is present in the graph otherwise it is discarded.

The ranking of user depends on the friend matching graph generated using the similarity measure. The ranking solely depends on graph structure of friend matching graph which contains two aspects: 1) how the edges are connected, 2) how much weight there is on every edge. Because, it states that a user having more edges has common lifestyle and that user can be recommended to most of the remaining. Let $N(i)$ denotes the set of neighbors of user i. Let $Rec = [Rec(1), Rec(2), ..., Rec(n)]^T$ denote the impact ranking vector where $Rec(i)$ is the impact ranking of user i in the friend-matching graph, and n is the number of users in the system. The calculation of $Rec(i)$ is defined as follows:

$$Rec(i) = \frac{\sum_{j=1}^{n} w(i, j).Rec(j)}{\sum_{j=1}^{n} w(i, j)}$$

where, w(i,j) = Similarity(i,j). The calculation of $Rec(i)$ is an iterative process because any change of its neighbours will change $Rec(i)$ accordingly. Therefore, we use a matrix representation to clearly get the iterative process.

3.3 Algorithm

Input: The query user x, The recommendation coefficient z (value lies between 0 and 1), The required number of recommended friends from the system n.
Output: Friend list Lx.

Steps:

```
(1)  Initialize Lx and Q as null
(2)  extracts x's life style vector Vx
(3)  for each lifestyle k the probability of which in vector Vx is not
     zero do
(4)      put users accordingly in the entry of k into Q
(5)  end for
(6)  for each user y not belonging to Q do
(7)      Similarity(x,y)=0
(8)  end for
(9)  for each user y in the database do
(10)     Rec x(y) = z*Similarity(x,y) + (1 - z)*Rank(y)*C
(11) end for
(12) sort all recommended users in decreasing order with respect
     to Rec x(y)
(13) put the top n users in the sorted list Lx
```

After extracting user's lifestyle from smart-phone, similarities between the users should be found out. In the above algorithm, Rec x(y) is the recommendation score of user y for the query user x, Similarity(x,y) is the similarity between user x and user y, and Rank(y) is the impact of user y. z lies between 0 and 1 is the recommendation coefficient characterizing user's preference. C is introduced to make Similarity(x,y) and Rank(y) in the same order of magnitude, which can be roughly set to u/10, where u is the number of users in the system. When z = 1, the recommendation is solely based on the similarity; when z = 0, the recommendation is solely based on the impact ranking.

4. ADVANTAGES

(1) If friends on social network share similar lifestyle, then recommend potential friends to users.

(2) The feedback mechanism allows us to measure the needs of users, by providing a user interface that allows the user to rate the friend list.

(3) Access to authorized person only which avoids duplication of accounts.

(4) Application provides privacy measures of user.

(5) System determine better recommendation based on lifestyle with high accuracy.

(6) It provides user friendly interface to user.

5. CONCLUSION

The existing system relies on social link analysis. 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. Thus, on the basis of literature survey and by analysing the existing system, we have come to a conclusion that the proposed system will not only suggest friends based on social graphs but the lifestyle of the user will also be taken into consideration. Development of this software will surely decrease the social and real-life friend recommendation gap. Future scope of this system is not limited. First, we would like to test the system on large scale which will eventually increase the accuracy of the system. And secondly, to work on a large scale, our system can be incorporated with existing social networking sites.

6. ACKNOWLEDGEMENT

We would like to extend my sincere gratitude and thanks to my guide Mr. A. M. Jagtap, for his invaluable guidance and for giving us useful inputs and encouragement time and again, which inspired us to work harder. Due to his forethought, appreciation of the work involved and continuous imparting of useful tips, this report has been successfully completed.

7. REFERENCES

[1] Z. Wang, J. Liao, Q.Cao and H.Qi. 'Friendbook: A Semantic-based Friend Recommendation System for Social Networks'. IEEE, pages 538 - 551, 2015.

[2] L. Bian and H. Holtzman. Online friend recommendation through personality matching and collaborative filtering. Proc. of UBI-COMM, pages 230-235, 2011.

[3] J. Kwon and S. Kim. Friend recommendation method using physical and social context. International Journal of Computer Science and Network Security, 10(11):116-120, 2010.

[4] X. Yu, A. Pan, L.-A. Tang, Z. Li, and J. Han. Geo-friends recommendation in gps-based cyber-physical social network. Proc. of ASONAM, pages 361-368, 2011.

[5] K. Farrahi and D. Gatica-Perez. Discovering Routines from Largescale Human Locations using Probabilistic Topic Models. ACM Transactions on Intelligent Systems and Technology (TIST), 2(1), 2011.

[6] K. Farrahi and D. Gatica-Perez. Probabilistic mining of socio-geographic routines from mobile phone data. Selected Topics in Signal Processing, IEEE Journal of, 4(4):746-755, 2010.

[7] T. Huynh, M. Fritz, and B. Schiel. Discovery of Activity Patterns using Topic Models. Proc. of UbiComp, 2008.