International Journal of Emerging Technology in Computer Science & Electronics (IJETCSE)
ISSN: 0976-1353 Volume 21 Issue 2 – APRIL 2016.

SEMANTIC BASED SUGGESTION SYSTEM FOR ACQUAINTANCES

B. Nikhila, D. Harshini Kalyan, Mary Vinothini, Sandeep Saraf

Department of computer science, SRM University
Department of computer science, SRM University
Department of computer science, SRM University
Department of computer science, SRM University

***Abstract*— Given the wide usage of various web services, there has been a rising need for a more accurate and an efficient recommendation system. The net usage of these web services has been contributed predominantly by various social networking sites which aim on building a platform to improve social relations among people who share similarities. These networking sites were initially aided with profile based friend recommendation systems and were further improved to include life style based activities. This was to improve the precision of suggesting a friend to the user based on ranking the similarities acquired from their daily activities rather than just matching profiles. We have updated this system by including features such as infograph construction and also automated blocking of previously blocked contacts.**

Index Terms – Recommendation System, Web Services, Life Style, Infograph.

I. INTRODUCTION

One challenge with existing social networking services is a way to advocate an honest friend to a user. Most of themsuppose pre-existing user relationships to choose friend candidates. For instance, Facebook depends on a social link analysis among people who already share common friends and recommends symmetrical users as potential friends. Sadly, this approach might not be the foremost applicable recent social science findings.

The rules to cluster individuals along include: 1) habits or life style; 2) attitudes; 3) tastes; 4) ethical standards; 5) economic level; and 6) individuals they already understand. Apparently, rule #3 and rule #6 area unit the thought -factors thought about by existing recommendation systems. Rule #1, though most likely the foremost intuitive, isn'twide used as a result of users' life designs area unit tough, if not not possible, to capture through internet actions. Rather, life styles are usually closely associated with daily routines and activities. Therefore, if we have a tendency to gather info on users' daily routines and activities, we will exploit rule #1 and suggest friends to individualssupported their similar life designs. Thus this disadvantage was overcome by incorporating life style based matching in addition to profile based friend recommendation. This new feature serves as a filter to recommend friends with whom we share much more similarities than just the profile and hence gives a more definitive reason to trust the suggestions made. Life style is based on various aspects in our everyday routine like places we visit, events we participate in, communities and groups we support and so on. Hence comparison based on these aspects maybe more efficient in suggesting a friend whom we can actually have useful relation with rather than strangers with common friends.

The suggested system hence requires collecting life styles of various users which are then stored in a cloud for future references. These data are collected using Latent Dirichlet Allocation. We can describe latent Dirichlet allocation (LDA) as a generative probabilistic model for collections of discrete data. LDA is a three-level hierarchical Bayesian model, in which each item of a collection is modeled as a finite mixture over an underlying set of topics. The goal is to find short descriptions of the members of a collection that enable efficientprocessing of large collections while preserving the essential statistical relationships that are useful for basic tasks such as classification, novelty detection, summarization, and similarity and relevancejudgments.

Latent Dirichlet allocation (LDA) is a generative probabilistic model of a corpus. The basic idea is that documents are represented as random mixtures over latent topics, where each topic is characterized by a distribution over words.

LDA assumes the following generative process for each document **w** in a corpus *D*:
1. Choose *N* _ Poisson(x).
2. Choose q _ Dir(a).
3. For each of the *N* words *wn*:
(a) Choose a topic *zn* _ Multinomial(q).
(b) Choose a word *wn*from *p*(*wn j zn;*b), a multinomial probability conditioned on the topic *zn*.

A *k*-dimensional Dirichlet random variable q can take values in the (*k*−1)-simplex (a *k*-vector
q lies in the (*k*−1)-simplex if q*i* _ 0, å*ki*=1 q*i* = 1), and has the following probability density on this
simplex:

$$p(\theta|\alpha) = \frac{\Gamma\left(\sum_{i=1}^{k}\alpha_i\right)}{\prod_{i=1}^{k}\Gamma(\alpha_i)}\theta_1^{\alpha_1-1}\cdots\theta_k^{\alpha_k-1}, \tag{1}$$

where the parameter a is a *k*-vector with components a*i*>0, and where G(*x*) is the Gamma function.

The Dirichlet is a convenient distribution on the simplex—it is in the exponential family, has finitedimensional sufficient statistics, and is conjugate to the multinomial distribution. In Section 5, theseproperties will facilitate the development of inference and parameter estimation algorithms for LDA.

Given the parameters a and b, the joint distribution of a topic mixture q, a set of *N* topics **z**, anda set of *N* words **w** is given by:

$$p(\theta,\mathbf{z},\mathbf{w}|\alpha,\beta) = p(\theta|\alpha)\prod_{n=1}^{N}p(z_n|\theta)p(w_n|z_n,\beta), \tag{2}$$

where*p*(*znj*q) is simply q*i* for the unique *i* such that *zin*= 1. Integrating over q and summing overz, we obtain the marginal distribution of a document:

$$p(\mathbf{w}|\alpha,\beta) = \int p(\theta|\alpha)\left(\prod_{n=1}^{N}\sum_{z_n}p(z_n|\theta)p(w_n|z_n,\beta)\right)d\theta. \tag{3}$$

Finally, taking the product of the marginal probabilities of single documents, we obtain the probability of a corpus:

$$p(D|\alpha,\beta) = \prod_{d=1}^{M}\int p(\theta_d|\alpha)\left(\prod_{n=1}^{N_d}\sum_{z_{dn}}p(z_{dn}|\theta_d)p(w_{dn}|z_{dn},\beta)\right)d\theta_d.$$

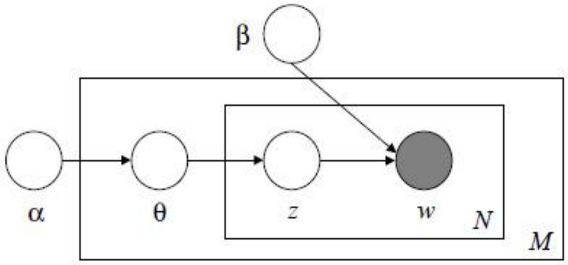


Figure 1: Graphical model representation of LDA.

The boxes are “plates” representing replicates.The outer plate represents documents, while the inner plate represents the repeated choiceof topics and words within a document.

Given the above features we have further improved the quality of friend suggestions by mainly updating two important features which avoid unnecessary friend recommendations and make matching much easier. The updates are as follows:

1. Infograph:

   Which gives a clear picture of the variouslife style based matching in the form of graph hence makes comparing the similarities easy which further help ranking of these friends based on these graphs.

2. Blocked contacts:

   Given the contacts which have already been blocked on the system working on and also the users known for bad reputations can be removed directly without suggestion. This thus makes the suggestions given more trust-worthy.

The rest of the paper is organized as follows. Section 2 covers related works. Section 3 provides the overall system architecture of the recommendation system. Section 4 provides a detailed explanation of the various algorithms and concepts used in this system. Finally, we conclude the paper and provide the future works in section 5.

II. RELATED WORKS

Recommendation systems are actually filtering systems, that predicts the rating or the preference that a user would give to an item. Recommendation systems are extremely common and are widely used in various applications such as search engines, social site tags, product queries in shopping sites, etc. Generally speaking, existing friend recommendation in social networking systems, e.g., Facebook, LinkedIn and Twitter, recommend friends to users if, according to their social relations, they share common friends.

Other recommendation systems have been proposed by researchers. Some of such systems are as follows:

1. Bian and Holtzman – MatchMaker – a collaborative filtering friend recommendation system based on personality matching.
2. Yu et al – a geographically related friends in social network by combining GPS information and social network structure.
3. Kwon and Kim – a friend recommendation method using physical and social context.

4. Cence Me –a recommendation system that usedmultiple sensors on the smartphone to capture user's activities, state, habits and surroundings.
5. SoundSense – a recommendation system that uses the microphone on the smartphone to recognize general sound types

The current recommendation systems are however different from the one proposed in this paper as this system utilizes the clustering algorithm to recommend friends based on their life style similarities. A similarity metric is used to measure the similarity between the life style of two users.

III. SYSTEM OVERVIEW

This section covers the system overview of the suggestion system. The Figure 2 shows the system architecture of the suggestion system, which assumes a client-server mode, where each client interacts with the server through data centers or cloud.

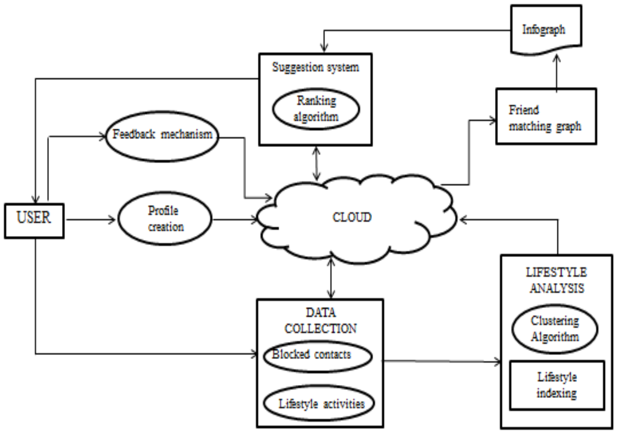


Figure 2: System architecture.

On client side, a 'user registration' module is created in which the user creates a profile and records his interests and likes. The profile is used to communicate with the cloud and record all of the user's day-to-day activities such as search pages on website, videos viewed and files downloaded, and clusters the data and lifestyle using LDA and clustering algorithm. As each user typically generates around 50MB of raw data every day, we choose MySQL as our low level data storage platform.

On the server side, we create four modules, designed to generate friend suggestions to various users. The 'data collection' module is responsible for gathering all the possible raw data about the user every day. This includes the various lifestyle information as well as syncing contacts and blocked contacts.The 'lifestyle analysis' module uses a clustering algorithm to cluster the various raw data, gathered in the previous module. Such clustering makes it easier to perform lifestyle indexing, which is the process of indexing or cataloging the clusters. The 'recommendation generation' module is used to generate the recommended friends for each user. This module transforms the data in clusters into a graph called infograph and uses a ranking algorithm on this graph to prioritize and rank the friends in accordance to their common interests. This module then provides the list of recommendations for each individual user. The 'feedback' module allows the user to either accept or reject or block the recommendations. It collects the response of each user and updates their profiles and grants access accordingly.

IV. SUGGESTION GENERATION

A. *k-MEANS CLUSTERING ALGORITHM*

For a given set of data points, clustering is the process of grouping a set of objects or data points in such a way that the objects or data points in one group are similar to each other than to those in other groups. Clustering algorithm is mainly used in data mining in many fields such as machine learning, pattern recognition, image analysis, etc.

Clustering can be formulated as a multi-objective optimization problem. It can be achieved by several different methods or algorithms. Some of the basic clustering models are connectivity model, centroid model, distribution model, density model, graph model, etc.

The algorithm used for this system is k-means clustering, which is a form of centroid model.

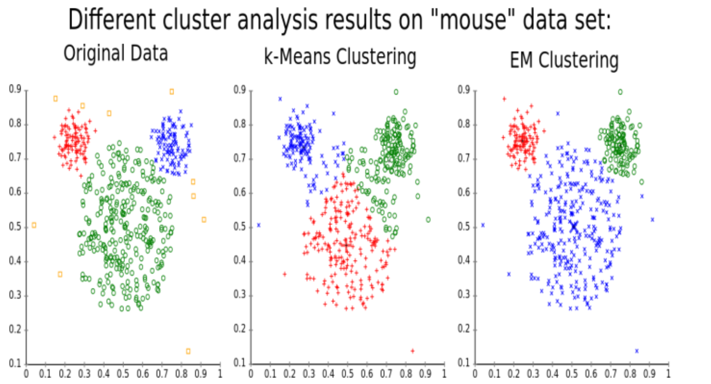
k-means clustering is a vector quantization, originally from signal processing. It is the most common and popular clustering algorithm used in data mining. This algorithm aims to partition n data elements in to k clusters. Each of the n data elements belongs to the cluster with the nearest mean. The k-means algorithm finds clusters with compatible spatial extent while the expectation maximization mechanism allows clusters to have different shapes.

Given a set of observations ($\mathbf{x}_1$, $\mathbf{x}_2$, …, $\mathbf{x}_n$), where each observation is ad-dimensional real vector, k-means clustering aims to partition the $n$observations into k ($\leq n$) sets $\mathbf{S} = \{S_1, S_2, …, S_k\}$ so as to minimize the within-cluster sum of squares (WCSS) (sum of distance functions of each point in the cluster to the K center).The objective is to find:

$$\underset{\mathbf{S}}{\arg\min} \sum_{i=1}^{k} \sum_{\mathbf{x}\in S_i} \left\| \mathbf{x} - \boldsymbol{\mu}_i \right\|^2$$

where $\boldsymbol{\mu}_i$ is the mean of points in S_i.

K-means clustering has the tendency to produce equi-sized clusters.



B. *FRIEND MATCHING GRAPH*

A friend matching graph is used to characterize the relationship between different users and pick out the similarity between the life style of the users. This system uses the concept of a similarity metric to measure the similarity to predict the likeliness that a user will like to know another user. The similarity metric used by this system is as follows.

Let Li = [p(z1|di), p(z2|di), ..., p(zZ|di)] and Lj = [p(z1|dj), p(z2|dj), ..., p(zZ|dj)] denote the life style vectors of user i and user j, respectively.

The similarity between user I and j are not affected by their life style vectors as a whole but an element within the vector, having the largest probability value. The similarity between these users, given by S(i, j), is defined as follows:

S(i, j) = Sc(i, j) · Sd(i, j)

whereSc(i, j) is used to measure the similarity of the life style vectors of users as a whole, Sd(i, j) is used to emphasize the similarity of users on their dominant life styles.

A friend matching graph is a weighted undirected graph G = (V, E, W), where V = {v1, v2, · · · ,vn} is the set of users and n is the number of users, E = {e(i, j)} is the set of links between users, and W : E → R is the set of weights of edges. There is an edge e(i, j) linking user i and user j if and only if their similarity S(i, j) ≥ Sthr, where Sthr is the predefined similarity threshold. The weight of that edge is represented by the similarity, that is, ω(i, j) = S(i, j).

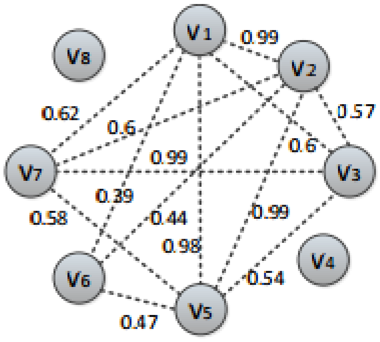


Figure 3.An example friend matching graph.

1.1 RANKING ALGORITHM

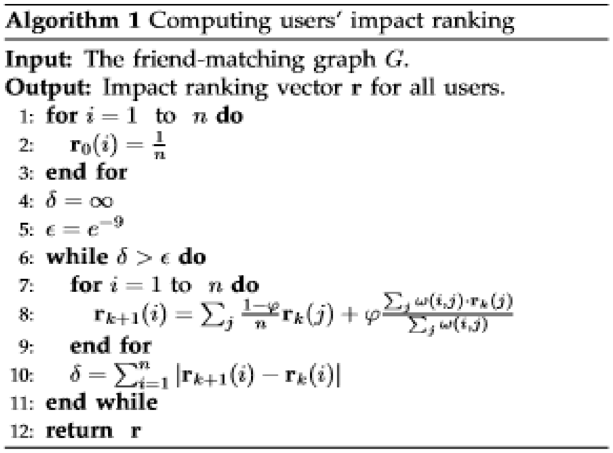
User ranking means a user’s capability to establish friendships in the network. In other words, the higher the ranking, the easier the user can be made friends with, because of their broader life styles. User ranking is developed from the concept of web ranking. Once the user ranking is obtained, it acts as a guideline to those who receive a recommendation list to choose friends.

The user ranking algorithm depends mainly on the structure of the friend matching graph. The two main aspects of the graph are the connected edges and the weight on every edge. Ranking should be used together with the similarity scores between the query user and the potential friend candidates, so that the recommended friends are those who not only share sufficient similarity with the query user, and are also popular ones through whom the query user can increase their own impact rankings.

Let N(i) denote the set of neighbors of user i. Let r = [r(1), r(2), ··· , r(n)]T denote the impact ranking vector where r(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 r(i) is defined as follows:

$$r(i) = \frac{\sum_{j \in N(i)} \omega(i,j) \cdot r(j)}{\sum_{j \in N(i)} \omega(i,j)}$$

A pseudo code for the user ranking algorithm is as follows.



```
Algorithm 1 Computing users' impact ranking
Input: The friend-matching graph $G$.
Output: Impact ranking vector r for all users.
for $i = 1$ to $n$ do
    $\mathbf{r}_0(i) = \frac{1}{n}$
end for
$\delta = \infty$
$\epsilon = e^{-9}$
while $\delta > \epsilon$ do
    for $i = 1$ to $n$ do
        $\mathbf{r}_{k+1}(i) = \sum_j \frac{1-\varphi}{n}\mathbf{r}_k(j) + \varphi\frac{\sum_j \omega(i,j)\cdot \mathbf{r}_k(j)}{\sum_j \omega(i,j)}$
    end for
    $\delta = \sum_{i=1}^{n} |\mathbf{r}_{k+1}(i) - \mathbf{r}_k(i)|$
end while
return r
```

V. CONCLUSION AND FUTURE WORKS

In this paper, we have presented the design and implementation of a semantic based suggestion system for acquaintances, which is different from the previously existing recommendations systems which work on the information solely provided by the user. The previously existing recommendation systems use LDA algorithm to collect and analyze raw data while this system utilizes clustering algorithm. This increases the efficiency and simplifies the algorithm further.

This system extracts the life style of various users based on their information and their day-to-day activities over the internet and generates friend recommendation for the user based on the similarity in their interests.

The current prototype, being advanced from the existing systems, could further be enhanced. The change in algorithm from LDA to clustering has proven to be more efficient and simpler. But we would like to implement lifestyle extraction using matrix vector multiplication in user impact ranking incrementally. This system can be enhanced further into a large scale field experiment. We also plan to implement life style extraction through more sensors and environmental reality platforms for more accurate and useful suggestions.

REFERENCES

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

[2] Friendbook - Zhibo Wang,Jilong Liao,Hairong Qi and Zhi Wang.

[3] Combating Friend Spam Using Social Rejection – QiangCao, Xiaowei Yang, Munagala. K.

[4] WMR a graph based algorithm for friend recommendation – Suchuan Lo, Chingching Lin

[5] An overview of secure friend matching in mobile social networks – Ganvir.R, Mahalle. V

[6] C. M. Bishop. Pattern recognition and machine learning. Springer New York, 2006.

[7] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent Dirichlet Allocation. Journal of Machine Learning Research, 3:993-1022, 2003.

[8] B. A. Frigyik, A. Kapila, and M. R. Gupta. Introduction to the dirichlet distribution and related processes. Department of Electrical Engineering, University of Washignton, UWEETR-2010-0006, 2010.

[9] L.Gou,F.You,J.Guo,L.Wu,andX.L.Zhang.Sfviz:Interestbased friends exploration and recommendation in social networks. Proc. of VINCI, page 15, 2011.

[10] G. Spaargaren and B. Van Vliet. Lifestyles, Consumption and the Environment: The Ecological Modernization of Domestic Consumption. Environmental Politics, 9(1):50-76, 2000.

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