ISSN 2047-3338

A Multifaceted Approach Towards Friend Recommendation in Social Network

S. Simranjit[1], M. Nikunj[2] and M. Nishant[3]
[1,3]Sardar Patel Institute of Technology, Mumbai, India
[2]Nirma Institute of Technology, Ahmedabad, India

***Abstract*— Social networking has become an intertwined part of our daily lives, with these websites having a user base of several hundred million. Friend recommendation system is a crucial aspect of these social networking platforms, but it hasn't received the importance it deserves. A good recommendation system would not only give the platform a more intuitive look but it will also improve performance of entire architecture. We have proposed a system using neural network and diversified weights based on multilayer text extraction, frequency of communication for friend recommendation from friends of friends. We have taken into account various factors which will assign score to his friends of friends and will recommend them more efficiently.**

***Index Terms*— Data Mining, Graph, Neural Network, Recommendation and Social Network**

I. Introduction

SOCIAL networks which started out as a platform for users to express their personality, has come a long way. The boom has propelled the business interests and hence academic attention in this domain. So, it is natural that in this era of Facebook, Google and LinkedIn recommendation engines in social networks have drawn upon interest among lot of researchers. Various recommendation engines are used for a various purposes like suggesting communities, advertisements and friend recommendation. Having the user instilling his faith in a recommendation engine can increase the probability of clicking on sponsor links.

Most recommendation systems singularly give emphasis to tags, explicit keywords provided by their users. However relevant these factors may be they are error prone and also incomplete. So in practice users are not getting accurate and desired recommendation. Added personalization text extraction with multiple weights of back propagation neural network will overcome this limitation. Here we have overcome this constraint and then integrated the most effective one with social information. This information is gathered by the reading histories, recommendations, user communication with others, personalities and likes from the active user's friends.

Network based approaches generally perform well in providing quality recommendations. Now a friend of user's friend rather than a random person will give highly desirable results. This approach implies a person is more likely to pursue a relationship based a common association. However, this does not provide any insights into human cognitive components, which is a multi-Dimensional belief system that may change over time. This approach still relies purely on the underlying structural properties of social networks. Since participants within social networks are humans, it would be of significant interest to approach the recommendation problem by supplementing network theory with cognitive theory.

II. Related Work

A. Community Detection

With the boom of the Internet, the web space of an individual has undergone a significant transformation to reflect the social life and characteristics. Owing to such massive migration, the Web now boasts of a very important component: Social Networking. The degree of communication affinity in this space has given rise to a tremendous volume of raw data that needs to be analyzed so as to create a better system and provide improved services. The research methodology of social network analysis is developed to understand the relationship between the various actors involved by studying and analyzing their communication affinity. The term actor refers to a person, an organization, an event or an object. Communication affinity can be in various forms depending on the networking service. In a social network, each actor is a node and many such nodes are connected by lines to depict relation between them. The social network structure graph is a graph that formed by those lines and nodes, and social network analysis is therefore a methodology that used to analyze the graph, and better understand the relationships among the actors in the social network so as to provide better services [1], [2], [3].

Different kinds of graph management and mining techniques are being studied, along with the corresponding applications. Note that the boundary between graph mining

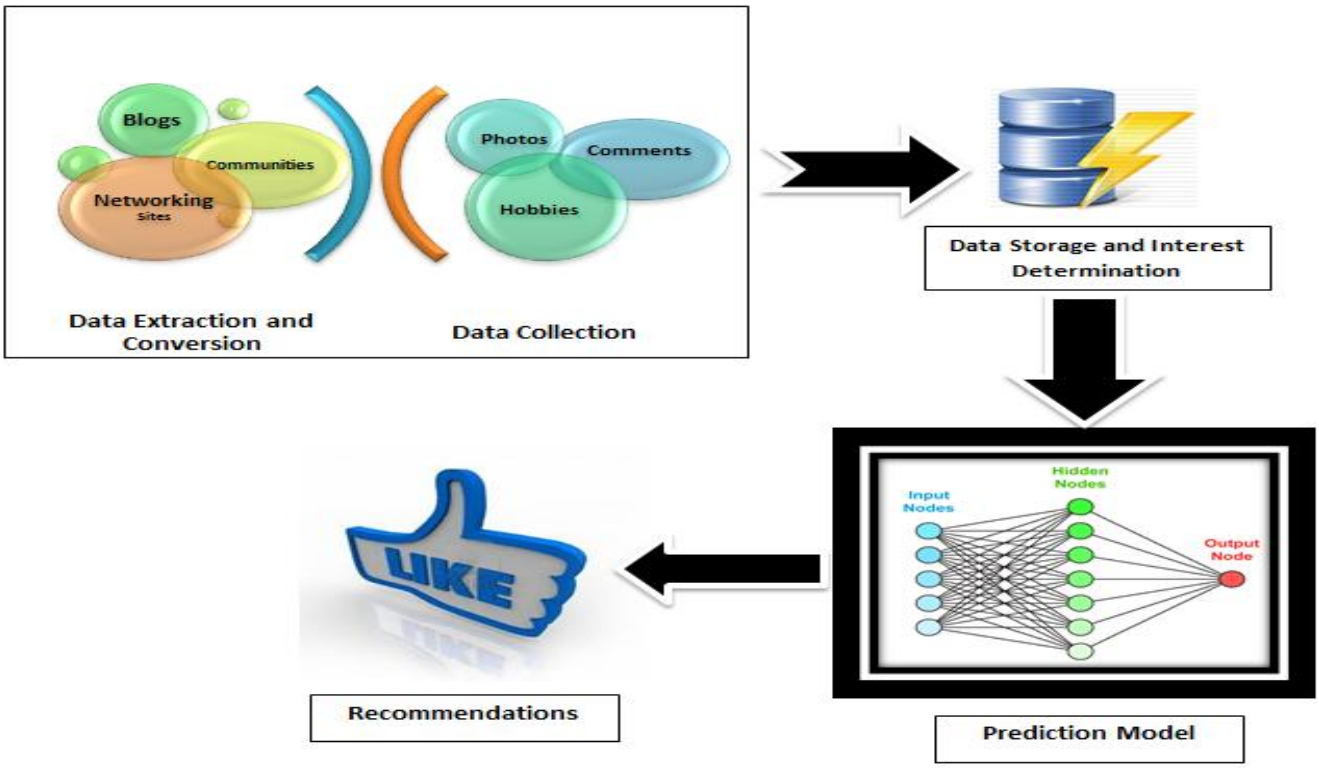

Fig. 1: Architecture of the recommendation system

and management algorithms is often not very clear, since many kinds of algorithms can often be classified as both. Safaei et al. [4] have proposed a way to analyze the relationship between a user and the communities it is a part of. This understanding equips us to recommend communities that might be of interest to the said user.

Provost et al. [5] have evaluated the concept of brand proximity to propose privacy-friendly methods for extracting quasi-social networks from browser behavior on user-generated content sites, for the purpose of finding good audiences for brand advertising (as opposed to click maximizing, for example). Nepusz and Bazso [6] have focused on the application of the maximum likelihood estimation in the case of graphs by presenting two stochastic graph models and two algorithms to fit them to datasets arising from real applications.

Chen et al. [7] have investigated the problem of mining frequent approximate patterns from a massive network by giving an approximation measure and show its impact on mining with the help of the gApprox algorithm. They have focused on how a pattern's support should be counted based on its approximate occurrences in the network. Zhang [8] et al. have proposed a method for identifying key users, based on mining of online social networks for marketing purpose by graph analysis.

B. Web Mining Techniques for understanding user preference

According to different analysis targets and resources, the web mining techniques can divided into three different types, which are Web Content Mining, Web Structure Mining and Web Usage Mining [9]. Web content mining is a web mining technique to analyze the content on the web. This content aside from text includes graphs, graphics, etc. [10]. Web content mining targets the knowledge discovery, in which the main objects are the traditional collections of text documents and, more recently, also the collections of multimedia documents such as images, videos, audios, which are embedded in or linked to the Web pages [11]. With improvements in bandwidth and storage the multimedia data embedded in web pages have proliferated. As a consequence images, audios, videos embedded in web pages are included as a part of web content [12]. In addition the websites providing users to comment on the content e.g. Blog posts or videos the natural language processing used is therefore the main technology that used in this area. The concept and techniques of Semantic Web and Ontology also have to be studied [13], [14].

Web structure mining is a technique that can be used to analyze the links and structure of websites [15]. Graph theory is usually the main concept and theory for web structure mining to analyze and explain the structure of websites. The different objects are linked in some way. Simply applying the traditional processes and assuming that the events are independent can lead to wrong conclusions. However, the appropriate handling of the links could lead to potential correlations, and then improve the predictive accuracy of the learned models [16]. In addition, the extraction of the structure of websites is always essential in this research area [17].

Web Usage Mining is the application of data Mining techniques to discover usage patterns from web data in order

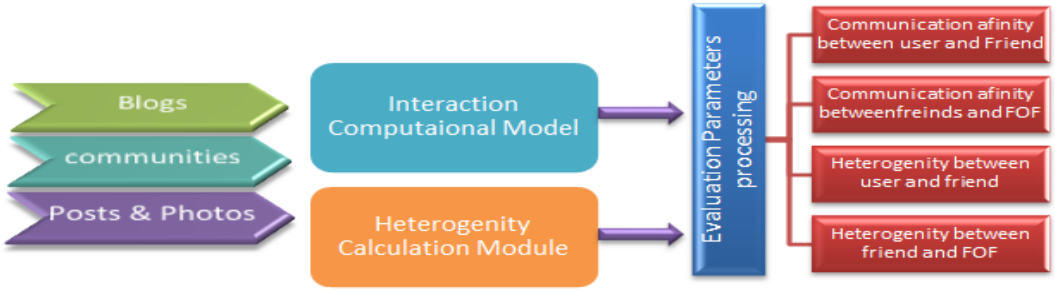

Fig. 2: Data collection, extraction with structured storage and filtering

to understand and better serve the needs of web based applications. In addition to the traditional web mining techniques user profiles should also be considered as a part of web mining [12].

III. Architecture

According to the aim and objective of this paper, we have designed system architecture to facilitate the development of the solution domain. The system will allow the multidimensional social data from blogs, communities, posts, tagging etc. to be collected. The social data will then be pre-processed by converting to structured data. This data will then be stored for retrieval during calculation phase. The architecture of our system is presented in Fig. 1.

A. Data Grooming

1) Data Collection and Extraction

This is the first step of the system. In this paper, we intend to collect all the social data which can be found by user's profile. The data may be in form of user's communication with others in the form of comments, photos, likes or personal details. One thing should be noticed that data collection process does not require the user to be logged in mandatorily. The users need not explicitly rate anything. The system itself will fetch the information from usage history. Since, users are not forced survey or give feedback unadulterated content can be obtained.

2) Conversion to structured data

The process of conversion to structured data is challenging. It itself has several steps such as purging which involves removal of data not conveying tangible information e.g. filtering out articles like the, a ,an. After the data has been purged it needs to be converted into form that can be used for actual calculation. This also might involve semantic analysis i.e. mapping multiple words having the same meaning to a single word or assigning domains to which that part of text might belong to. This structured data helps in identifying user related information and assigning initial weights. Moreover, structured data improves clustering coefficient. It also helps to know initial topic of interests of the user.

3) Data Storage

After data collection and extraction, the output of the data extraction will the data will be stored in a database. The database is designed according to the characteristics of different sources of social data.

B. Communication affinity Calculation

In this module a directed network graph is constructed for the user and edges are created which represent the score of communication between the nodes. We have described the calculation in much depth in Section IV.

C. User Interest Determination

In this module structured data is mined and initial user interest is determined. The detail of how the data will be processed and extracted will be discussed in later section of the paper. Hence, in this stage the sample space is narrowed down by eliminating some potential bad recommendation. So the components having minute value can be minimized precisely.

D. Prediction Module

This is the phase in which calculations take place. The user data obtained from the previous modules is fed to the neural network. The neural network generates a recommendation list based on the data and user feedback through friend selection.

IV. Units

In the network users are represented as nodes and edges between them represent their communication affinity. The weight of the edges describes the affinity of the user for his friend. If there is an edge from node a to node b we can automatically infer that there is an edge from node b to node a, because if a person appears in the users friend lists then vice-versa is also true. But, the graph is a directed graph so, the weight of edge from node "a" to node "b" is not necessarily equal to weight of edge from node b to node a. Weight of the edges describes the proximity of the users. All the people who are friends with users' friends but are not friends with the user are typically known as friends of friends or FOFs.

A. Communication affinity

It represents communication between node a and node b

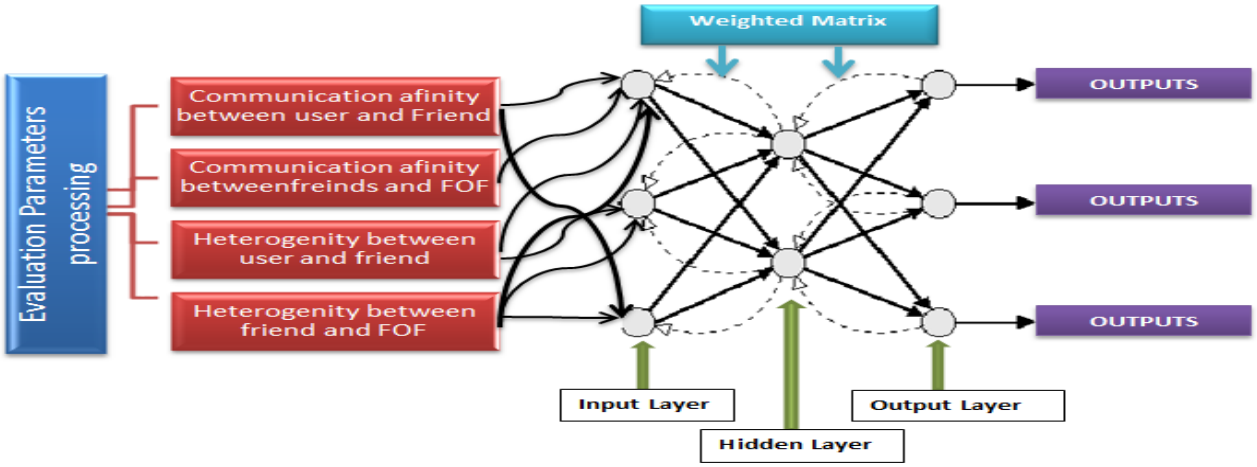

Fig. 3: BPNN Model using Evaluation Parameters as Input

done by node a i.e., how much node a has communicated with b. It is difficult to represent because many different and multifold methods of communication on a social networking website. If we consider Facebook which provides features such as comments likes tagging, blogs, messages and chatting communication affinity between a and b can be considered as a function of all of these features, where a's communication affinity with b will consist of tagging b, liking b's posts, blogs, pictures sending him messages, chatting with him etc. Its internal implementation and assigning weights to different elements is variable since it depends on the social network platform and also has security considerations. Here we have chosen to represent Communication Affinity of for a user as the percentage score of total of Communication Affinity for the user.

CA (a, b) is the communication affinity score of "a" with "b". (1)

B. *Topics of Interests*

This represents the topics which fascinate the user. The topics can be of very wide range and depth. This can be of a bigger scope like sports, politics, art, sciences or can be narrower and much more focused like basketball, Russian politics, modern European art and theory of relativity. The topics of interest can be extracted from mining the text of whatever the user writes on the network or from the communities he joins. Moreover Topics that characterize a given knowledge domain are somehow associated with each other. Those topics may also be related to topics of other domains. Hence, documents may contain information that is relevant to different domains to some degree. Hence, the use of fuzzy sets to represent user interests is an appropriate choice as it semantically captures the user's choice.

Let there exist 3 topics sports (t1), politics (t2) and art (t3). So for example

(TOI) user = [{t1, 0.44}, {t2, 0.37}, {t3, 0.53}] (2)

It will denote 44% participation in sport 37% in politics and 53% in art.

C. *Difference of interest*

It represents how aloof two users are in terms of different domains they are interested in. This is a highly abstract term and is not easily expressed. Here we shall define DOI between 2 users as the Euclidean distance of topics of interests.

Therefore:

$$DOI(a,b) = \sqrt{(a_1 - b_1)^2 + (a_2 - b_2)^2 + \cdots + (a_n - b_n)^2} \quad (3)$$

Thus TOI for a and b are as mentioned bellow.

TOI of a= [{t1, 0.1}, {t2, 0.37}, {t3, 0.53}]
TOI of b= [{t1, 0.4}, {t2, 0.1}, {t3, 0.5}]
DOI (a, b) =0.4047

Since DOI gives how aloof the users are from each other its reciprocal gives us similarity of interests. So similarity index

D. *Equation for scoring a FOF*

Thus the score of a FOF can depend on following factors: Communication affinity of user with all mutual friends of FOF and the user, Communication affinity of mutual friends with FOF and Difference of behaviors of user and FOF.

Thus the equation can be represented as:

$$S(a,b) = \frac{\sum_{i=1}^{i=n}[CA(a,M_i)^{\alpha} \times CA(M_i,b)^{\beta}]}{DOI(a,b)^{\gamma}} \quad (4)$$

Where,
M is the set of common friends between user and FOF
Mi represents ith member of common friend set
α is the user dependent weight assigned to the score of CA between user and friend.

β is the user dependent weight assigned to the score of CA between mutual friend and FOF.γ is the user dependent weight assigned to the Difference of interest between user and FOF.
Based upon the above equation we assign scores to the FOF of the users.

E. Determination of weights

Since, each user has his criterion for friend selection the values of weights (α, β and γ) for different users will be different. So, to determine weights of different parameters we use simple feedback neural network.

This model uses various back propagation neural networks (BPNN) as shown in Fig. 3. BPNN use a supervised learning mechanism, and are constructed from simple computational units referred to as neurons. Neurons are connected by weighted links that allow for communication of values. When a neuron's signal is transmitted, it is transmitted along all of the links that diverge from it. These signals terminate at the incoming connections with the other neurons in the network.

In a BPNN, learning is initiated with the presentation of a training set to the network. The network generates an output pattern, and compares this output pattern with the expected result. If an error is observed, the weightings associated with the links between neurons are adjusted to reduce this error. The learning algorithm utilized has two stages. The first of these stages is when the training input pattern is presented to the network input layer. The network propagates the input pattern from layer to layer until the output layer results are generated. Then, if the results differ from the expected, an error is calculated, and then transmitted backwards through the network to the input layer. It is during this process that the values for the weights are adjusted to reduce the error encountered. This mechanism is repeated until a terminating condition is achieved.

The characteristics, preference and social behaviors vary dramatically among human beings. Neural network-based recommendation mechanism is special for its leaning and forecasting ability to imply the implicit relationships behind these factors and requester's pattern of preference. Notably, a forecasted score for each FOF will be obtained and the three weights will be learned through the neural network, with the user selection acting as feedback for correction of weights. The iterative process of recommendation and selection will train the network enabling it to generate more accurate results.

V. CONCLUSION

In this paper we have proposed a system for recommending friends of friends which can be used widely in many social networking applications. The breakthrough point of this system is its application of BPNN method for scoring friend of friend and using fuzzy sets to represent user interests in various topics. These multiple weighted values with multiple iterative neural network output make the system more robust and reliable. Our method of breaking down behavior into interaction and topics of interests are good candidates for giving semantically accurate results. This will give an apt statistical snapshot of human behavior in the social network. The snapshot of this behavior can be used by social networking platforms for giving better recommendations to the user and has commercial benefits for all stakeholders. Quantification of human behavior and recommendation engines are open ended fields with a lot of scope. Moreover text mining and natural language processing have their own challenges like the web having its own jargon which keeps on evolving. Hence capturing the essence of rich data available and it's usage in these networking platforms is still an uphill task.

REFERENCES

[1] Freeman, L, "Centrality in Social Networks: Conceptual Clarification," *Social Networks*, vol. 1, pp. 219-239, 1979.

[2] *Social Network Analysis: A Hand Book* 2nd ed., SAGE Publications Ltd., J. Scott 2000.

[3] S. Wasserman and K. Faustb *Social Network Analysis: Methods and Applications*. New York: Cambridge University Press 1994.

[4] Marjaneh Safaei, Merve Sahan, Mustafa Ilkan, "Social Graph Generation & Forecasting using Social Network Mining", in *Proc. 33rd Annual IEEE International Computer Software and Applications Conference*, Seattle, 2009, pp. 31-35.

[5] Foster Provost, Brian Dalessandro, Rod Hook, Xiaohan Zhang, Alan Murray, "Audience Selection for On-line Brand Advertising: Privacy-friendly Social Network Targeting", in *Proc. 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, Paris, 2009, pp.707-716.

[6] Tam´as Nepusz, "Maximum Likelihood Methods for Data Mining in Datasets Represented by Graphs", in *Proc. 5th International Symposium on Intelligent Systems and Informatics*, Subotica, 2007, pp. 161-165.

[7] Chen Chen† Xifeng Yan‡ Feida Zhu† Jiawei Han†,"gApprox: Mining Frequent Approximate Patterns from a Massive Network", in *Proc. 7th IEEE International Conference on Data Mining*, Omaha, 28-31 Oct. 2007, pp.445-450.

[8] Yu Zhang, Zhaoqing Wang, Chaolun Xia, "Identifying Key Users for Targeted Marketing by Mining Online Social Network", in *Proc. 24th IEEE International Conference on Advanced Information Networking and Applications Workshops*, Perth, 2010, pp. 644 - 649.

[9] Ting, I. H. "Web Mining Techniques for On-line Social Networks Analysis" in *Proc. 5th International Conference on Service Systems and Service Management*, Australia, 2008, pp. 696-700

[10] Agrawal, R., Rajagopalan, S., Srikant, R., and Xu, Y. "Mining Newsgroup Using Networks Arising From Social Behavior" in *Proc. 12th International World Wide Web Conference*, Budapest, 2003, pp. 529-535

[11] da Costa, M.G., Jr., Zhiguo Gong "Web structure mining: an introduction", in Proc. *IEEE International Conference on Information Acquisition*, Hong Kong and Macau, 2005, pp. 590-595.

[12] Jaideep Srivastava, Robert Cooley, Mukund Deshpande, PangNing Tan, "Webusage mining , discovery and applications of usage patterns from Web data," *ACM SIGKDD Explorations Newsletter*, vol. 1, pp. 12-23, June 2000.

[13] Godbole, N., Srinivasaiah, M., Skiena, S, " Large-Scale Sentiment Analysis for News and Blogs" in *Proc. ICWSM*, Boulder Colorado, 2007, pp. 219-222

[14] Mika P, "Flink: Semantic Web Technology for the Extraction and Analysis of Social Networks," *Journal of Web Semantics*, vol. 3 pp. 211-223, October 2005.

[15] Dingt C. H. Q., Zha, H., Husbands, P., and Simont, "H. D. Link Analysis: Hubs and Authorities on the World Wide Web," *SIAM Review*, vol. 46, pp. 256-268, June 2004. missing.

[16] L. Getoor, "Link Mining :a new data mining Challenge" *ACM SIGKDD Explorations*, vol. 4, pp.84-89, July 2003.

[17] Fu, F., Chen, X., Liu, L., and Wang, L, "Social Dilemmas in An Online Social Network: The Structure and Evolution of Cooperation" *Physics Letters A*, vol. 371, pp. 58-64, November 2007.