International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 6.887
Volume 6 Issue V, May 2018- Available at www.ijraset.com

Hybrid Friend Recommendation Approach based on Clustering and Similarity Index

Sumit Kumar Sharma[1] Manisha Lokhande[2]
[1]*Research Scholar Sri Aurobindo Institute of Technology, Indore*
[2]*Assistant Professor Assistant Professor*

Abstract: ***With the change in internet, way of using it is also changing. Internet not only just provide a way of interaction through operating system but also involving in different fields like artificial intelligence, machine learning etc. Social networking site is mostly used and popular platform of internet. It creates connectivity among people with similar features. Internet service in social networking is making it essential in life of people. Keeping in mind about user nature and interest makes it easy to recommend similar characteristic friend. It also provides a way of enhancing business, promoting products, getting current new, updates etc. It helps to be in touch with our contacts by recommending them similar characteristics of user. Here, a hybrid recommendation model has been proposed and developed to explore the similarity between users based on lifestyle basis. It is the combination of K-mean Clustering and Similarity Weight Calculation to explore the more precise and absolute results. The complete solution is developed using Java technology and evaluated on basis of precision, recall and f-score.***

Keywords: Recommender systems; Friend Recommendation; K-mean; Clustering approach; Filtering approach

I. INTRODUCTION

Data mining, the extraction of hidden prognosticative data from massive databases, could be a powerful new technology with nice potential to assist corporations concentrate on the foremost vital data in their knowledge warehouses. Data processing tools predict future trends and behaviors, permitting businesses to form proactive, knowledge-driven selections. The machine-controlled, prospective analyses offered by data processing move on the far side the analyses of past events provided by retrospective tools typical of call support systems. Data processing tools will answer business queries that historically were too time overwhelming to resolve. They scour databases for hidden patterns, finding prognosticative data that consultants might miss as a result of it lies outside their expectations.

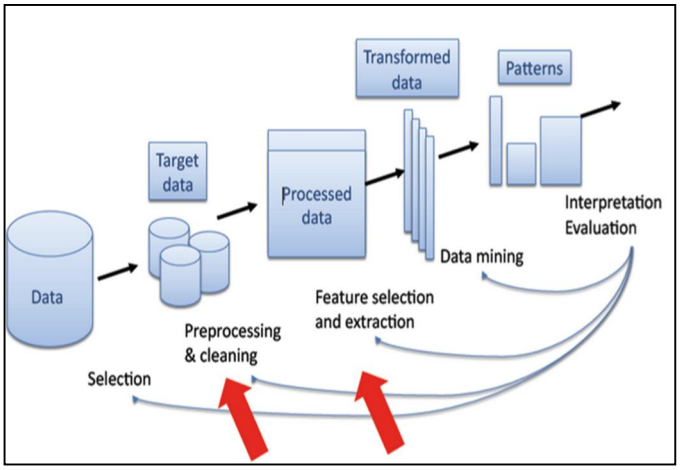


Figure 1.1: Data Mining for recommendation

Kacchi & Deorankar et al. In[1] address that lifestyle is one of the major and important part of social networking websites. To address their need and overcome the existing solution, they developed a solution model based on filtering and recommendation system. Proposed solution is based on data collection and analysis method with friend matching graph and ranking steps. The

proposed system will work like a client-server application where the user which is requesting the query acts as a client. Life documents of each user are collected from the client with the help of browser. In this phase, these collected data will be stored into a file either in semi-structured or structured format accordingly. The life styles of users are extracted by using either Hadoop technology or SQL depending on the type of file as input to it. Then the concept of reverse indexing is used for easy retrieval of the desired data. Then with the help of graph data structure we can represent the relationship between users. As recommendation is based on different priorities like similar interest, similar blood group, nearby location, ranking is also one of the factor. So, the ranks of users are calculated using the pseudocodes mentioned in this paper. Finally, client/user sends a query and server will respond a list of friends to the user/client (browser) accordingly. A block representation of proposed solution is shown in figure 2.1.

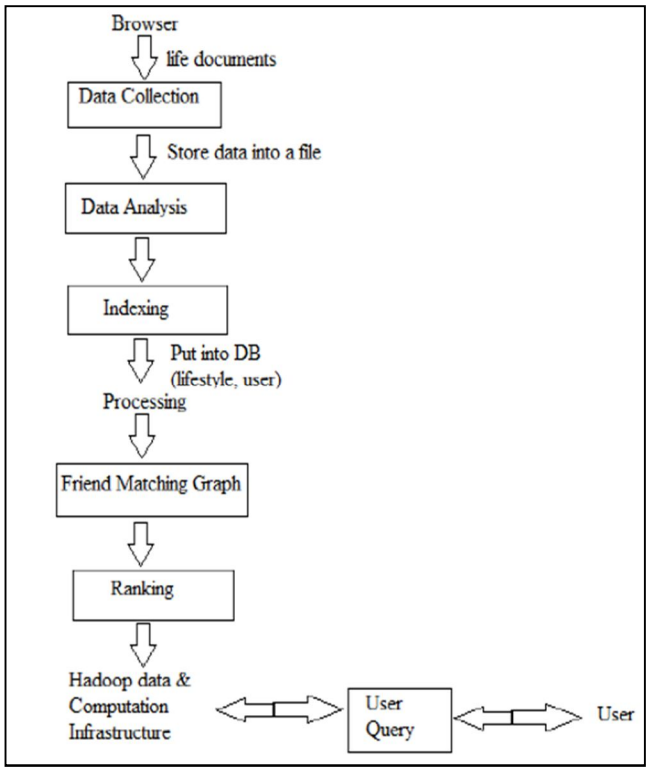


Figure 2: Friend Recommendation System

Machine learning [ML] is kind of artificial intelligence that provides ability to learn and take decision with intelligence system. It concentrates to develop computer program having capability to change with respect to variation in data. Subsequently, data mining [DM] is an approach to extract knowledge and characteristics of data based on features. The only difference between ML and DM is ML focused on prediction approach based on learning from training data whereas DM focused on analysis approach and knowledge discovery from unknown properties of data.

Wang et al. In[2] address that machine learning and data mining approach can be used as recommendation purpose and can effectively perform the role of friend recommender on social networking sites. Here, they extract user features and activity as the input source of recommendation and perform semantic analysis approach for recommendation. It only concentrate to current activity and can't quantify the complete nature of user. In social networking applications users attempt to connect with friends and other user based on common nature and area of interest. Few works consider this philosophy and attempt to extract to find friend based on closeness with other user profile.

Kwon et al. In[3] finds that social networking users always prefer recommendation system due to heavy volume of friend information. They consider context aware algorithms to extract user nature based on physical and social context. They attempt to propose hybrid approach based on common score of physical and context features. They also emphasize to use of recommendation algorithm for more accurate results.

Bian et al. In[4] has developed an solution based on personality of users and uses collaborative filtering for recommendation purpose. This study concludes that personal information not only reflects the user nature but also help to obtain its interest rate. This can be helpful to extract friend group nature and recommendation purpose. For example if user share same native place and same

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 6.887
Volume 6 Issue V, May 2018- Available at www.ijraset.com

school during childhood time, they may have common interest behind friendship. User matching network and personality extraction can be an good way to find relevance among social networking users.

II. PROBLEM DOMAIN

The information overloading and irrelevant information extraction is major problem of today. Information portals and renowned sources consists large quantity of data uploaded for various Subjects. Hectic schedule and poor knowledge of technology exhaust the user during information searching and information retrieval. It becomes pathetic when user recently joins any social networking sites. Initially, user account start with zero friend and user may find it worthless and boring. Here, finding the old friend from huge user list and not be possible to send friend request. Subsequently, user may want to connect across the world people who match their lifestyle and can help in their problem. Lifestyle and activity are the two most common factors which can be used for friend recommendation. It may help to meet user searching and effective friend connection.

Everyone has difference perceptions and different reading liking. It may vary as per user preference and job requirement. Popularity of content and impact of information is also important for user search. Exploring the particular lifestyle from the user is the essential phenomena to recommend relevant friends. This problem becomes more sensitive and crucial when we try to extract current affairs and Friends from large online sources.

A Friends suggestion can be explored from various sections like city, sports, editorial, international, national, entertainment etc. All this sections have equal importance and different user followers. Some time there may be possibility that, they may consist relevant information but in different sections and different Friends papers. Friends Recommendation System can overcome this problem and suggest relevant Friends according to user preference and popularity factor. A lifestyle based friend recommendation has been suggested by few authors which are discussed in related work section. The complete study concludes that there is need to develop popularity and uses based recommendation tool to gather popular and important relevant Friends at one place.

III. METHODOLOGY

Clustering is an approach to classify all elements in such a way that every similar element should be resided into single group based on their similarity. Subsequently, it also reside irrelevant elements into another group based on their similarity value and maximum cluster size. Here, K-

Mean clustering approach has been used to construct group of similar users based on lifestyle similarity. It is one of the simplest unsupervised learning algorithms which simplify the work of mining by classifying the similar elements in cluster using k-centroids parameter. It calculates distance between each element to evaluate similarity and reside them into single cluster by comparing with k-centroid parameter.

The most significant challenge in this section was to map the lifestyle of each user and convert them into quantify figures. Here, quantification technique has been implemented to convert all similar values into matching score. A snipping of this work is shown in figure 3.1

Afterwards, output has been forwarded to clustering module for cluster making. This complete phenomenon generates the most relevant users based on similarity distance.

A. Recommendation Approach

Recommendation system is a method of filtering that use rating, similarity score or preference score to predict the frequency between item and elements. Recommendation systems have become increasingly popular in recent years due to wide area of applications and use into movies, books, articles etc. prediction and suggestions. Here, a customized recommendation algorithm has been developed and append with K-mean clustering algorithm to provide more accurate and relevant solution. Here, recommended cluster is used as input data source and similarity score has been calculated.

Similarity score represents the total lifestyle closeness of each user with desire user lifestyle. In simple words, high similarity score represents more lifestyle closeness in comparison with users having low similarity score. Similarity threshold value has been used to filter out the retrieved user id and recommend most close user references. High threshold would represent filter with high strength and it will recommend users with most close values. Subsequently, low threshold represents low filter strength and high numbers of users. The complete phenomena of proposed solution have been shown in figure 3.3

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 6.887
Volume 6 Issue V, May 2018- Available at www.ijraset.com

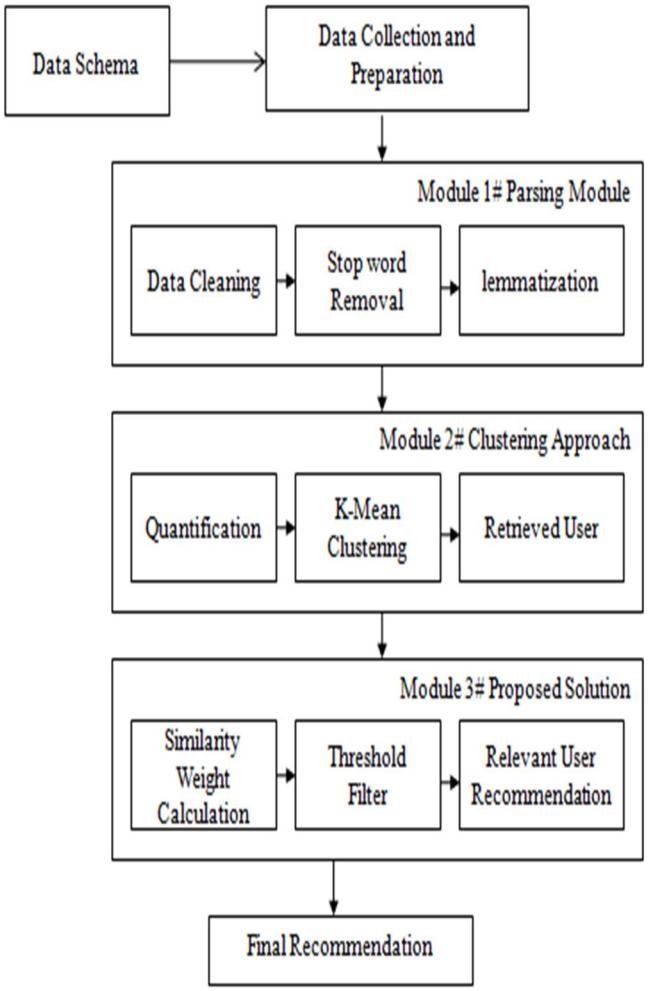


Figure 3.3: Proposed Architecture

The complete system can be understood by following points:

B. Data Schema

A self design schema has been used for data collection and analysis purpose. Initially 43 [Forty Three] attributes are considered for data collection purpose which consist Varity of questions along with different options for answer. Here, some of objectives kind and some are subjective based questions.

C. Data Collection and Preparation

A real data collection process has been performed to draw more accurate and absolute results. A survey of around 350 people has been made using googledocs and prepared for further process.

D. Modules

Afterwards, complete solution is further classified into three modules, can be listed as below

1) Module 1: This module is design to implement dedicated parser of proposed solution. Here, a filtration process has been integrated to prepare more accurate and precise results.

a) Data Cleaning: This method is intended to remove the unwanted tuples from inserted database based on incomplete and irrelevant transactions.

b) Stop Word Removal: It removes the stop words in the subjective answer section to simplify the quantification process.

c) *Lemmatization:* Analyzing words with vocabulary and morphology lemmitization aims at removing words which does not belongs to dictionary and includes only words of dictionary and is called as Lemma. If confronted with the token *saw*, stemming might return just *s*, whereas lemmatization would attempt to return either *see* or *saw* depending on whether the use of the token was as a verb or a noun. This module has been implemented to simplify the analysis of subjective answers.

2) *Module 2:* This module is designed for clustering purpose. A K-Mean clustering section has been suggested in it.
   a) *Quantification:* It is approximation of a subjective aspect (attribute, characteristic, property) of a thing or phenomenon into numbers through an arbitrary scale. Every aspect of nature can be quantified although it may not be measurable. This module has been implemented to convert subjective answers into number form based on similarity factor.
   b) *K-Mean Clustering Algorithm:* A K-mean Clustering algorithm has been used **to** prepare set of similar tuples.
   c) *Retrieved User: It* accommodates all relevant users for requested user id. It considered all ID of current cluster where requested user lies. So It precise the recommendation based on similarity.

IV. RESULT ANALYSIS

A java based recommendation tool has been developed to implement the proposed solution. Proposed implementation view has been classified into four modules. Initially, all collected data has been exported into.CSV file format from Google docs and loaded to perform parsing process. Incomplete data removal technique has been implemented for data cleaning purpose. Lemmatization and tokenization are used for the removal of unwanted words and these and they come in the categories of Stanford library. The module used in this project provides with clustering process for more accurate and exact data source. The major challenge during clustering process was a mapping of all user lifestyle into a numeric representation. A self-developed quantification process has been used to convert all content sentiments into numeric figures. Subsequently, K-mean clustering approach has been performed to extract similar users based on desired user information. No doubt clustering can help us to retrieve relevant users but can't be considered as recommendation technique. A self-proposed recommendation technique based collaborative filtering has been implemented for suggestion purpose. Here, similarity weight has been considered to evaluate the ranking of a user in similarity index cluster. Similarity weight is the total sum of all weights estimated during quantification process. At last threshold value has been used to filter out all retrieved document and generate most relevant users as a final recommendation. Recall-precision and F-score parameters have been used to measure the performance of proposed solution. The complete performance has been evaluated on basis of Recall [Accuracy], Precision and F-Score [Final Score]. A view of selected users ID and threshold values are shown in

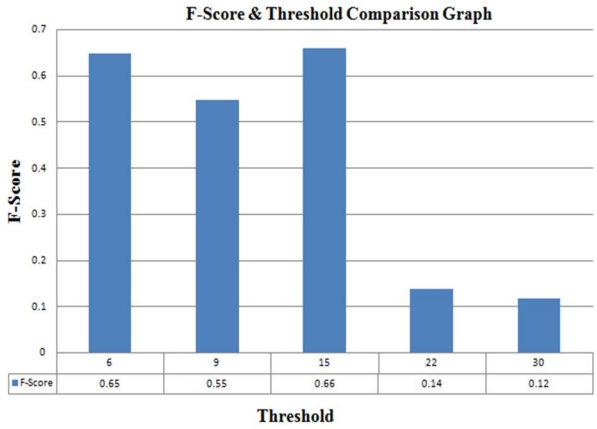
Table 1(a) & Table 1(b) for Cluster size 2.

S.No.	User ID
1	1
2	6
3	8
4	32
5	64
6	78
7	81
8	100
9	106
10	115

Table 1(a): List of All selected User ID

Threshold	Strength
6	Very Low
9	Low
15	Medium
22	High
30	Very High

Table 1(b): List of All selected Threshold



V. CONCLUSION

The complete work concludes that a lifestyle based friend recommendation system can add a boost up feature into normal intranet communication or social networking sites. In this work, a modified version of clustering and filtering approach has been proposed as the hybrid solution for recommendation purpose. Subsequently, a custom quantification and filtering approach have been performed to simplify the recommendation process with accurate performance. A java based recommendation tool has been developed and evaluated the performance on basis of recall, precision, and f-score. Here, certain observations has been recorded which are listed below.

A. Constant recall with value one (1) has been recorded for 6 and 9 thresholds.
B. into threshold value decrees the system performance and its lacks gradually with respect to enhancement of threshold value.
C. The variable precision score has been recorded with for different users. Maximum 0.86 precision and minimum 0.09 score has been recorded
D. An integrated F-Score has been calculated for each threshold value where minimum 0.12 and maximum 0.66 scores have been recorde
E. The complete work concludes that proposed solution can be used as the recommendation technique for friend recommendation purpose
F. A hybrid recommendation model has been proposed and developed to explore the similarity between users based on lifestyle basi
G. It is the combination of K-mean Clustering and Similarity Weight Calculation to explore the more precise and absolute results.

REFERENCES

[1] Bahram Amini, Roliana Ibrahim, Mohd Shahizan Othman, “Discovering The Impact Of Knowledge In Recommender Systems: A Comparative Study”. International Journal of Computer Science & Engineering Survey (IJCSES) Vol.2, No.3, August 201
[2] G. Adomavicius and A. Tuzhilin. Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions. IEEE TKDE, 17(6):734–749, 2005

[3] Neeraj raheja, v.k.katiyar, "survey on ameliorate data extraction in web mining by Clustering the web log data" International Journal of Computer Science Issues" Vol. 11, pp-2, 2014.
[4] Minsuk Kahng, Sangkeun Lee, Sang-goo Lee, "Ranking in Context-Aware Recommender Systems". Proceedings of the 20th International Conference on World Wide Web, WWW 2011, India, March 28 - April 1, 2011
[5] Ch.Nagini, M.Srinivasa Rao, Dr. R.v.krishnaiah, "An implementation view for news recommendation system" international journal of engineering research & technology, vol. 2,pp-701-704, 2013.
[6] Tofik R. Kacchi, Anil V. Deorankar, "Friend recommendation system based on lifestyles of users". 2Nd International Conference on Advances in Electrical, Electronics, Information, Communications and Bio-informatics (AEEICB), 11 August 2016
[7] 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
[8] J. Kwon and S. Kim, "Friend recommendation method using physical and social context," Int. J. Comput. Sci. Netw. Security, vol. 10, no. 11, pp. 116–120, 2010
[9] 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. 230–235.
[10] G. Linden, B. Smith, and J. York, Amazon.com Recommendations: Item-to-item Collaborative Filtering, IEEE Internet Computing, 7(1), pp. 76-80, 200
[11] Kanungo, T, Piatko, Mount, D. M Netanyahu, N.S Piatko, C. D.; Silverman, R.; Wu, A. Y. "An Efficient K-mean Clustering Algorithm: Analysis and implementation"(PDF). IEEE Trans. Pattern Analysis and Machine Intelligence. 24 (7): 881–892. Retrieved 2009-04-24.