International Journal on Recent and Innovation Trends in Computing and Communication ISSN: 2321-8169
Volume: 3 Issue: 8 5263 - 5267

Enforcing recommendation in Social networks By user interests

Shruti Agrawal
Computer Engineering Department
SKN Sinhgad Institute of Technology and Science
Lonavala, Pune, India
shruti1591@gmail.com

Prof. Thombre V. D
Computer Engineering Department
SKN Sinhgad Institute of Technology and Science
Lonavala, Pune, India
vdt.sknsits@sinhgad.edu

***Abstract*—** It is a human tendency to get others opinion before doing something, this we can see much more in friends circle. As the internet spreads over the world it interferes much more in our daily life, for example applications like Facebook and whatsup becomes the inseparable part of internet savvy peoples. So it is obvious that most of the users search their friends based on their taste or keywords using the recommendation system. Many recommendation systems are existed which provides the recommendation by capturing users taste or profile data in the social networking site. This paper put forwards an idea of creating recommendation system based on the collected user comment data from the social networking site pages using an efficient web crawler. This method enhances to get the recommendation from many social networking sites in a given instance. This makes the system as an independent adaptive model which can be easily apply on many social networking sites to get user recommendation for the given query. System strongly empowered by the well grained NLP protocols with fuzzy classification approach.

Keywords : *Web crawler, NLP, fuzzy logic, recommendation, web parsing.*

*****

I. INTRODUCTION

Recommender systems are the systems which suggest the users about relevant product, items, and user interest based on the relevant data. Because of the advent of social networking sites such as Facebook, twitter more and more users are trying to give there reviews, ratings on their interested topics. Because of the increasing nature of the users, huge amount of data is get collected. So it will get difficult for the recommendation system to read the data and give recommendation.

A recent survey conducted at USA universities shows that 25 %of product selling is increased due to the recommendation systems. More than 90 % of people are agreed on the things that are recommended by their friends and 50 % of the users buy the recommended things. So recommendation plays very important role in business applications, social circle and in various related areas.

Figure 1 shows the precise difference between the traditional search system and famous recommendation systems.

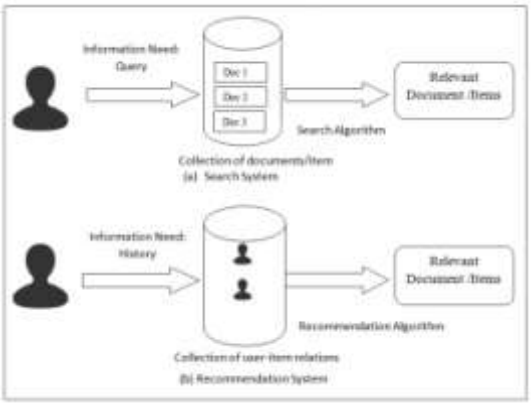

Figure 1: Difference between searching and recommendation

Preprocessing is a well-known technique of reducing data size for faster computation. In data mining and machine learning techniques a lots of unused data is presents along with the useful data , so that reduces the efficiency of the task processing. Often these unwanted data misleads the results, so in such applications preprocessing is at core part. Generalized preprocessing has four steps as below.

1. Data Cleaning
2. Data Integration
3. Data transformation
4. Data reduction

- Data cleaning: It is a technique of filling the missing data, smoothing the noisy data and removing the outliers. Data cleaning helps a lot in finding the attributes of interest.

- Data Integration: Data Integration is a technique of gathering the data from multiple stores at a same place so it will get easy to retrieve the data from the same source.

- Data Transformation: Data Transformation is a technique putting the data in more appropriate form so it can be used effectively in mining process. Data transformation uses different sub techniques such as normalization, smoothing, aggregation, generalization etc.

- Data Reduction: In this technique complex datasets are minimized to its simpler forms without comprising the originality of the data. Stemming is one of the best techniques especially used for the purpose of the data reduction. In this technique root word of derived word is find out in such way that meaning of the word will not change much. Like Stemming there is one more method known as lemmatization which goes in parallel with the stemming but with the slight difference. In case of stemming a set of rules are applied on

derived words but here part of the speech is not considered at all. In contrast in lemmatization part of speech and the meaning of the word is first understood and then root word is obtained.

Feature extraction is a normal process that comes in data mining; normally it is done to reduce the dimensions by selecting only those parts of the data that leads to the result. Numbers of methods are proposed to extract the features from the massive amount of datasets. Some of the techniques are described below.

1. Principal Component Analysis
2. Linear Discriminant Analysis

Feature extraction process is normally carried out in four important steps i.e. generate subset, evaluate subset, stop condition and validate the result.

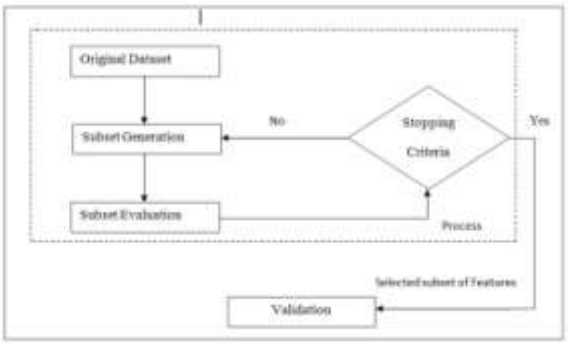

Figure 2: Feature extraction process

- Subset generation: In this process a certain strategy is applied to generate the subset of features from the original dataset.
- Subset Evaluation: Here generated subsets are tested and evaluated against the evaluating criterion. The criterion is set to obtained the good quality of the subsets.
- Stopping criterion: Here a criterion is set to stop feature extraction process. [1] Gives the entire possible stopping criterion that can be used.

Result validation: Once feature extraction process stops, the generated subsets are validated by using the prior knowledge about the data.

The rest of the paper is organized as follows. Section 2 discusses some related work and section 3 presents the design of our approach. The details of the results and some discussions we have conducted on this approach are presented in section 4 as Results and Discussions. Sections 5 provide hints of some extension of our approach as future work and conclusion.

II. LITERATURE SURVEY

This section represents all the related works of technologies used in our proposed model.

[2] Focused on all the described data preprocessing technique. This paper elaborates the operations of each of these techniques and also its sub techniques in prescribe manner. [3] Focuses on stemming techniques, paper gives a good difference between the stemming and lemmatization, advantage of one over another and sub techniques that can be used under main techniques.

[4] Presents a deep survey on various dimensionality reduction techniques that are used more often. The author wrote about how the problem of feature extraction can be solved easily by taking two methods i.e. PCA and Linear Discriminant Analysis. While doing this research authors also described the various statistical measures such as information theory, Mutual Information, Information Gain (IG), Gain Ratio (GR), Symmetric Uncertainty (SU), Correlation- Based Feature Selection, Statistics (CHI) etc.

Fuzzy logic is technique in which nonlinear mapping of an input data to a set of output data is calculated. Now a day's fuzzy logic gets lots of attention as it emerging as a one of the best technique to get answer when problem has a diverse behavior. Fuzzy logic has a wide application area such as information technology, analyzing the data, decision making, and pattern recognition. Fuzzy logic has four main parts such as The Fuzzifier, A Rule Base, An Inference Engine and A Defuzzifier.

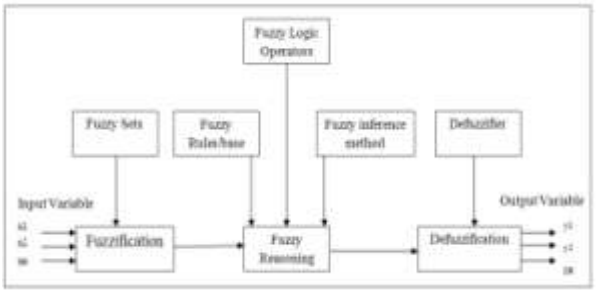

Figure 3: Fuzzy logic steps

Recommendation systems are widely classifies in three categories as Collaborative Filtering, Content Based and Hybrid Recommendations.

[5] Elaborates fuzzy logic based recommender system which makes use of triangular fuzzy number. To do recommendation author makes use of E commerce domain as this domain still facing problems for recommendation. Author uses Fuzzy near Compactness technique to find the similarity between the needs of the consumer and features of the product. To proves the effectiveness of the system publishers makes use of 50 different laptops from Sony or Lenovo. Now a day's most of the recommendation systems make use of collaborative filtering to give recommendation about the product but the collaborative filtering is not well suited for the one to one item recommendations like events.

[6] Describes the one and one item recommendation which is again based on the fuzzy logic as it improves the efficiency of the system. An author proves the importance of the algorithm in various domains such as E-commerce, E-Learning, and E-Government. As a future work of the paper authors trying to implement the algorithm for the realistic applications such as trade exhibition recommender system.

Disambiguation, less preciseness, incomplete nature of resultant data is some of the normal occurring problems of the recommendation systems. So to reduce this problem [7]

proposed a one theory which is based on the Fuzzy Theoretic Method (FTM). This method is capable to handle the item that has random probabilistic nature.

FTM uses representation method to extract the features of the item and feedback on those items, also it uses various similarity measures of fuzzy logic such as Jaccard Index, Cosine, Proximity or Correlation similarity measures, and recommendation strategies such as the maximum-minimum or weighted-sum fuzzy theoretic recommendation strategies. Movie dataset has been used to show the experimental evaluation of the system against the existing systems. Finally authors conclude that due to the lower model and recommendation size system provides a good accuracy over others.

[8] Implements an E-election system based on the fuzzy recommendation. Main aim of the system is to help the voters about the candidate that are nearby to the voter's preferences and the voter's identity to enhance the participation of voters. This algorithm is best suited for the scenario where events are going to be happens only once.

Also publisher's uses fuzzy clustering graph which shows the similarity between the different political parties to the citizens so it will get easy for citizens to select the proper candidate. The system is intended to give competition to the smart vote system which is pretty famous. Although they are in competition, the techniques used by these systems are different. In smart vote system similarities are find out by using "Match Point" while above system finds the similarities best on the distances of the high dimensional spaces.

III. PROPOSED METHODOLOGY

The idea of this proposed method is triggered by the fact that online social networking site users are often getting very bad recommendation for their searched results. This is due to old techniques or old posted comments of the user on the pages. So we are proposing a technique of providing a fresh recommendation for the users, for their searched keyword on the online social networking sites.

For this we are creating enriched online social networking site that will run in the LAN, Where users allowed posting the comments. On the other hand our system will recommend the right users for their fresh posts on the web page.

In this section, we describe our approach of Dynamic Recommendation for social networking sites with heuristic approach for the steps shown in figure 4.

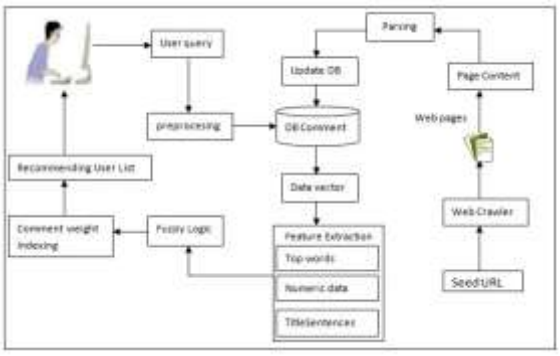

Figure 4: Over view of proposed approach

Step 1: In this step we are creating a web crawler which accepts a seed URL of all users of social networking site and searches it's all links.

Web crawlers are an essential component to search engines; running a web crawler is a challenging task. There are tricky performance and reliability issues and even more importantly, there are social issues. Crawling is the most fragile application since it involves interacting with hundreds of thousands of web servers and various name servers, which are all beyond the control of the system.

Web crawling speed is governed not only by the speed of one's own Internet connection, but also by the speed of the sites that are to be crawled. Especially if one is a crawling site from multiple servers, the total crawling time can be significantly reduced, if many downloads are done in parallel.

Despite the numerous applications for Web crawlers, at the core they are all fundamentally the same.

Web crawler in our system fetches the web page data and parses them to free from html tags to identify the user comments and store in the database in recursive manner for a assigned time. The below algorithm is used for the deployment of the web crawler.

Algorithm 1: DFS(G, v)

Input: Seed URL
Output: All SUB URL's of the site
Step 0: Start
Step 1: Enter into the URL and read the web page content
Step 2: Parse the web page content
Step 3: Identify sub URL and mark the current URL
Step 4: repeat step (1, 2, and 3) for the entire sub URL's
Stop 5: Stop

Algorithm 1: Depth first Algorithm

Step 2: Here user enters a query to search desired friends over the social networking sites.

Step 3: In this step preprocessing is carried out by the system which comprises of three sub parts: special symbol remover, stop word identification and word stemmer.

- ✓ Special symbol removal : here all the special symbols of the input data is removed as these symbols are not having any contribution in result generation

- ✓ Stop words: stop words are the supporting words used to add more meanings to the data. To accomplish this task a dictionary is used which contains a predefined stop words. On removing of these words raw meaning of the data is not changing at all. So stopword removing catalyses the process of mining technique more vigoursly.

- ✓ Word stemmer: it is a process of converting the derived words to its base form.

Step 4:-Feature extraction: Feature extraction plays very important role in semantic data analyzing. Here in this system we make use of three vital features for the precise answer extraction, viz title sentence, numerical data and term weight.

- **Title sentence**

Title sentence plays very important role when it used as a feature for the extraction purpose. The reason behind this is they provide the important narration about the documents. Very first statement of the comment is always used as a title sentence for that comment. Aside from the above use, they can be used to give convenient name to that document.

- **Numerical data.**

Numbers plays important role in recommendation area. So it becomes a crucial to find out such numerical data.

- **Term weight**

Term weight refers to a step of finding the most repeated words, as these words are the word that represents the most of the sematic of the documents. Hence our system identifies such words for more clarity in the recommendation.

Step 5: *Fuzzy Logic* - In this step fuzzy logic is imposed on extracted features to have precise output. The fuzzy logic is carried out in four prescribed function.

- First function took all the extracted features of the feature extraction steps as a fuzzy crisp values.
- Once fuzzification is done its output is fed to the fuzzy inference engine where if then rule is applied on the fuzzified values.

When the data is in final stage of fuzzy, its score is finding out by setting filtering protocols to get the set of user names as the recommendation.

The feature extraction can be done using fuzzy logic based on following equation

$$f(x) = \int_0^1 \sum_1^n (T_i, T_s, N_d)$$

Where
T_i= Topword Detection
T_s= Title Sentence
N_d= Noun Detection
$F_{(x)}$ =feature Summarized set

Step 6: Recommendation – After getting summary all summary words are compared with the query words to get the weight of the summary with respect to the query. Then system will recommend the users whose weight is more than or equal to 1. This process shown in the below algorithm

Recombination can be done by using following equation

$$R(x) = \int_1^n f_i$$

Where
f_i =feature Summarized Words
Q=Query
R(x)=Recommendation

The complete process of recommendation can be represent by the following pseudo code

OVERALL SYSTEM PSEUDEO CODE

Input : Set of comments C_i and user query Q
Output : Related Comments R_i
Step 0 : Start
Step 1 : Fetch all comments C_i
Step 2 : for i=0 to C_i length
Step 3 : Preprocess comments
Step 4 : find numeric data
Step 5 : Find top words
Step 6 : end for
Step 7 : For query Q find title word T_w
Step 8 : for i=0 to C_i length
Step 9 : Find Numeric score N_s , Top Word Score Tw_s , Title Score T_s
Step 10 : Add Ci,N_s,TW_s,T_s to vector V
Step 10 : end for
Step 11 : for i=0 to V length
Step 12 : for each V_i find fuzzy score
Step 13 : find sum of fuzzy score of individual V_i
Step 14 : Sort V_i in descending order according to sum
Step 15 : For user query Q check its occurrence in V_i to find query related comment R_i;
Step 16 : return R_i

IV. RESULTS AND DISCUSSIONS

To show the effectiveness of proposed system some experiments are conducted on java based windows machine using Apache tomcat as the server. To measure the performance of the system we set the bench mark on different number of query words for different run of recommendation. And then we allow the number of users to seek the recommendation from the system. To evaluate the performance of the system MAE parameter is considered.

In statistics, the mean absolute error (MAE) is an entity which is used to measure how close forecasts or predictions are to the eventual outcomes. The mean absolute error is given by

$$\mathrm{MAE} = \frac{1}{n}\sum_{i=1}^{n}|f_i - y_i| = \frac{1}{n}\sum_{i=1}^{n}|e_i|.$$

As the name suggests, the mean absolute error is an average of the absolute errors $|e_i| = |f_i - y_i|$, where f_i is the prediction and y_i the true value. Note that alternative formulations may include relative frequencies as weight factors.

On observing MAE for our model which is powered with fuzzy logic with the personalized Recommendation model mentioned in [9], Then the accumulated result can be shown in the below table no 1.

Sr No	Personalized Recommedation Model (PRM)	Proposed Model
1	0.9	0.6
2	0.9	0.8
3	0.8	0.8
4	0.8	0.6

Table 1 : Mean Absolute error for the PRM And proposed system

The graph plotting of both the models can be seen in below figure 5.

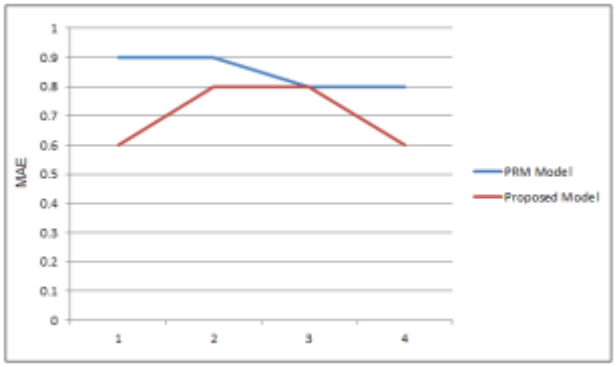

Figure 5 : MAE comparison of two models

Figure 5 graph indicates that proposed system is having less MAE value than the PRM model .This shows the better performance of our idea of using web crawler with enriched NLP protocols for the recommendation system.

V. CONCLUSION AND FEATURE SCOPE

The proposed system successfully designs a recursive multi-threaded web crawler which actually takes a seed URL from the online social networking site and crawl it thoroughly to collect its entire sub URL's. Then another baby crawler in the system crawls each and every collected web page to get the parsed information of the each web page. Proposed system extracts the important features (like title sentences, most repeated words and numerical data etc.)From the user comments or posts using strong NLP protocols. These feature scores are using as crisp values for fuzzy logic to classify the summary for the recommendation. Then finally by using similarity measure between comment summary and user query respective users are recommended by maintaining high accuracy.

The proposed system can be enhancing as an effective API that can be easily integrate with any social networking sites. This can be done by some parameter settings like

- ✓ Setting Seed url of the site
- ✓ Setting Social networking sites API integration for Jason object for https protocol
- ✓ Adding Social networking site licenses

REFERENCES

[1] Peng Liu, Naijun Wu, Jiaxian Zhu, Junjie Yin, and Wei Zhang, "A Unified Strategy of Feature Selection",The Second International Conference on Advanced Data Mining and Applications(ADML 2006), China, August 2006, pp. 457 – 464.

[2] " A Comprehensive Approach Towards Data Preprocessing Techniques & Association Rules", Jasdeep Singh Malik, Prachi Goyal, Mr.Akhilesh K Sharma Assistant Professor, IES-IPS Academy, Rajendra Nagar Indore – 452012 , India

[3] "A Comparative Study of Stemming Algorithms " Ms. Anjali Ganesh Jivani , Anjali Ganesh Jivani et al, Int. J. Comp. Tech. Appl., Vol 2 (6), 1930-1938

[4] "A Survey on Dimensionality Reduction Technique " V. Arul Kumar1, N. Elavarasan2 *International Journal of Emerging Trends & Technology in Computer Science (IJETTCS)*

[5] "A Fuzzy Logic Based Personalized Recommender System" Ojokoh, B. A., Omisore, M. O, Samuel, O. W, and Ogunniyi, T. O. *IRACST - International Journal of Computer Science and Information Technology & Security (IJCSITS), ISSN: 2249-9555Vol. 2, No.5, October 2012*

[6] "One-and-only item recommendation with fuzzy logic techniques" Chris Cornelis a,*, Jie Lu b, Xuetao Guo b, Guanquang Zhang b Information Sciences 177 (2007) 4906–4921

[7] "Fuzzy Modeling for Item Recommender Systems Or A Fuzzy Theoretic Method for Recommender Systems " Azene Zenebe, Anthony F. Norcio

[8] "A Fuzzy Recommender System for eElections" Luis Ter´an and Andreas Meier , Information Systems Research Group, University of Fribourg.

[9] Xueming Qian, Member,He Feng, Guoshuai Zhao and Tao Mei , "Personalized Recommendation Combining User Interest and Social Circle " , IEEE TRANSACTIONS KNOWLEDGE AND DATA ENGINNERING VOL:26 NO:7 YEAR 2014