

Introducing Recommendation as Per Users Choice In Social Networks

Shruti Agrawal

Information Technology Department
Vidyalankar Institute of Technology

Email ID-editorcassstudies@gmail.com

Abstract— It is a human nature to get others opinion on social networking before doing something. As the internet demand spreads over the world it is interfering much more in our daily life, for example applications like Facebook and whatsapp becomes a very important part of internet savvy peoples. So that most of the users on internet search their friends based on their taste or keywords using the recommendation system. Many recommendation systems already exist which provides the recommendation by capturing users taste or profile data in the social networking site. This paper puts an idea of creating recommendation system based on collected user comment data from the social networking site page using an efficient web crawler. This method increases to get the recommendation from many social networking sites in a given instance. This makes existing system as an independent adaptive model which can be easily apply on many social networking sites to get user recommendation for the given query. Final system strongly empowered by NLP protocols with fuzzy classification approach.

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

I. INTRODUCTION

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

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

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

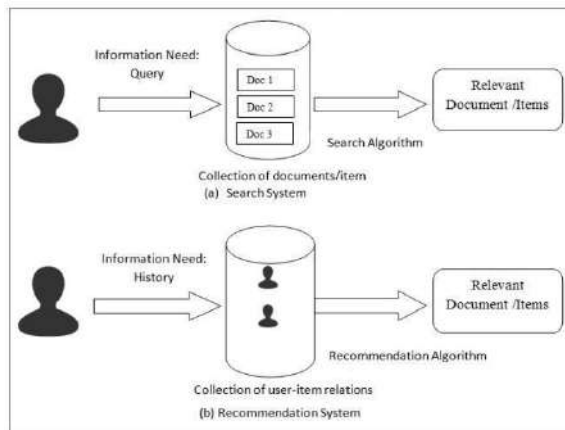



Figure 1: Difference between traditional searching and recommendation

Preprocessing is a well-known technique of reducing data size for faster computation. In data mining and machine learning techniques lot of unused data is presents along with the useful data , so that reduces the efficiency of the task processing. Often these unwanted data misleads many 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 finding and filling the missing data, smoothing the noisy data and removing the outliers. Data cleaning helps a lot in finding the final attributes of interest.
- **Data Integration:** Data Integration is a technique of gathering the data from multiple places and 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 of putting the data in more appropriate form so it can be used effectively in mining process. Data transformation uses different sub techniques like normalization, smoothing, aggregation, generalization etc.

Access this Article Online	Quick Response Code: 
Website: http://heb-nic.in/view-latest-issue	
Received on 22/12/2018	
Accepted on 25/12/2018 © HEB All rights reserved	

- **Data Reduction:** In this technique the complex datasets are minimized to its simpler forms without comprising the originality of the data. Stemming is one of the best techniques especially to be used for the purpose of the data reduction. In this technique root word of derived word is find out so that meaning of the word will not change much. Like Stemming there is one more method known as lemmatization which works 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 is included in data mining; normally it is done to reduce the dimensions by selecting only those parts of the data which leads to the result. Numbers of methods are proposed to extract the feature from the massive amount of datasets. Some of the techniques are described below as:-

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.

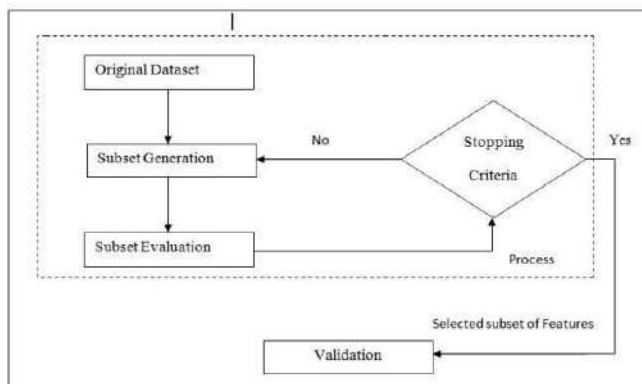


Figure 2: Feature extraction process

- **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

- [1] This section represents all related works of technologies used in our proposed model.
- [2] Focused on all the described data preprocessing technique. This paper elaborates many operations of each of these techniques and also its sub techniques in prescribe manner.
- [3] Focuses on the 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 the main techniques.
- [4] Presents a very 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 the 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 the 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 in all existing techniques 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 major parts such as The Fuzzifier, A Rule Base, An Inference Engine and A Defuzzifier.

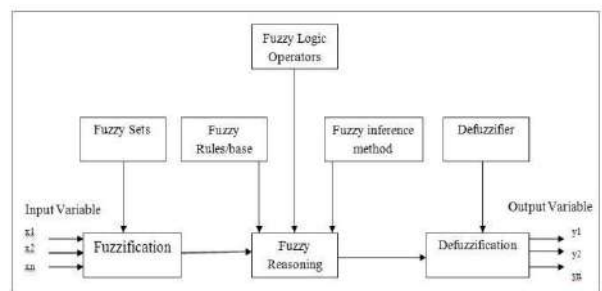


Figure 3: Fuzzy logic steps

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

[5] Elaborates fuzzy logic based recommender system that 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 out the

similarity between the needs of the consumer and the features of the product. To prove the effectiveness of the system publishers make 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 on the product but the collaborative filtering is not well suited for the one to one item recommendations like events.

[6] Describes the one and only one item recommendation which is again based on the fuzzy logic as it improves the efficiency of the system. The 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 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 the 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 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 the 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 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 to post 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 steps shown in figure 4.

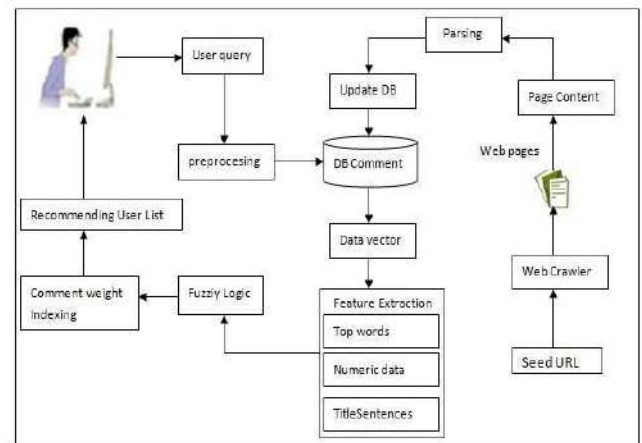


Figure 4: Over view of proposed approach

Step 1: In this step we are creating a web crawler which accept 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 many 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 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 have 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 of Web crawlers, at the core they are all fundamentally the same.

Web crawler in our system fetches the web page data and parses it 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: graph G and a start vertex v of G

Output: labeling of edges of G in the connected component of v as discovery edges and back edges set Label (v, VISITED)

```

for all e ∈ G.incidentEdges(v)
if getLabel(e) = UNEXPLORE
w ← opposite(v,e)
if getLabel(w) = UNEXPLORE
setLabel(e, DISCOVERY)
DFS(G, w)
else
set Label(e, BACK)

```

Algorithm 1: Depth first Algorithm

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

Step 3: This is the step where we are preprocessing is conducted, where string is processed to its basic meaning word by the following four main activities: Sentence Segmentation, Tokenization, Removing Stop Word, and Word Stemming.

- *Sentence segmentation* is the boundary detection and separating source text into sentence.
- *Tokenization* is actually separating the input query into individual words.
- *Stop word removal* :In any document narration the conjunction words does not play much of the role in the meaning of the document, so by discarding these words (like: is, the, for, an) from the documents which greatly reduces the overhead of processing
- *Stemming*: Many of the elongated words in the English language generally fail to provide the proper meaning in the given scenario and also they increases the computational time. So it is necessary to bring the words to their base form by replacing its extended characters with desired characters (Example: studied to study, where ied is replaced with y).

Step 4:-Feature extraction: As mentioned in earlier segments the features of a text play a very great role in the semantic categorization techniques. So our system makes use of four different features, which are extracted from the parsed data of the web pages which are stored in the database by using different techniques as mentioned below.

- *Title sentence*

In any text the title sentences are actually providing the most of abstract of the narration. So the extraction of title sentence is contributing an important role in identifying the summary of the text. System considers a very first sentence of the text as the title sentence. Another use of title sentence is to assign a proper name to the outcome result.

- *Numerical data.*

The numbers in any narration greatly affects the quality of document. So the system identifies the numbers and extract from the text to form a numerical vector.

- *Term weight.*

The most repetitive words in text are obviously the important words. So system identifies the list of the most repeated words and considers some top n elements (where n is user defined) as the important word for text to store in vector.

Step 5: Fuzzy Logic - The aim of text summarization is based on the extraction method of sentence selection. One of the methods to get the appropriate sentences is to assign some of numerical measure of a sentence for the Summary known as sentence weighting and then select the best ones. Therefore, the features score of each sentence is that we termed in the prior section are used to acquire the significant sentences. In this section, we use method to extract the essential sentences: text Summarization is based on fuzzy logic method. The system involves of the following core Steps:

Step A: In the fuzzifier, crisp inputs are taken, which are result of the feature extraction.

Step B: After fuzzification, the inference engine refers to the rule base containing fuzzy IFTHEN rules.

Step C: In the last step, we get the final sentence score. In the inference engine, the most important part is the definition of fuzzy IF-THEN rules. The essential sentences are extracted from these rules according to our features criteria. Sample of IF-THEN rules are described below.

IF (NoWordInTitle > 0.81) and (SentenceLength > 0.81) and (TermFreq > 0.81) and (SentencePosition > 0.81) and (SentenceSimilarity > 0.81) and (NoProperNoun > 0.81) and (NoThematicWord > 0.81) and (NumericalData > 0.81) THEN (Sentence is important).

After this process all the text sentences are ranked in descending order according to their scores. A set of uppermost score sentences are extracted as a text summary.

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= Topwords Detection

T_s= Title Sentences

N_d= Noun Detection

F_(x) =feature Summarized set

Step 6: Recommendation – After getting summary all the 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 comment C_i and user query Q
 Output : Related Comment R_i

- Step 0 : Start
- Step 1 : Fetch all the comments C_i
- Step 2 : for i=0 to C_i length
- Step 3 : Preprocess the comment
- Step 4 : find numeric data
- Step 5 : Find out top words
- Step 6 : end of for
- Step 7 : For query Q find out title word T_w
- Step 8 : for i=0 to C_i length
- Step 9 : Find Numeric score N_s , Top Word Score T_ws , Title Score T_s
- Step 10 : Add C_i,N_s,T_ws,T_s to the vector V
- Step 10 : end of for loop
- Step 11 : for i=0 to V length
- Step 12 : for each V_i find fuzzy score
- Step 13 : find the sum of fuzzy score of individual V_i
- Step 14 : Sort V_i in descending order according to the sum
- Step 15 : For user query Q check its occurrence in V_i to find out 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 server. To measure the performance of the system we can set the bench mark on different number of query words for different run of recommendation. And then we will 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 the forecasts or predictions are to the eventual outcomes. The mean absolute error is given by

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

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 the alternative formulations may include relative frequencies as weight factors.

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

Sr No	Personalized Recommendation 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

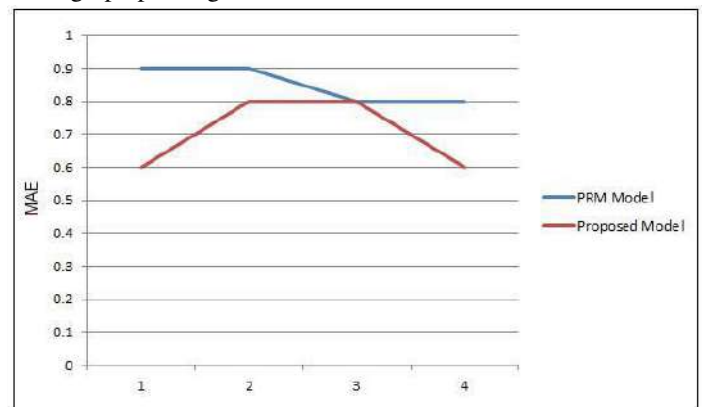


figure 5.

Figure 5 : MAE comparison of two models
 Figure 5 graph indicates that proposed system which is having less MAE value than the PRM

model .This shows the better performance of proposed idea of using web crawler with enriched NLP protocols for the recommendation system.

V. CONCLUSION AND FUTURE SCOPE

The proposed system successfully designs a recursive and multi-threaded web crawler which actually takes a seed URL from the online social networking site and crawls it thoroughly to collect its entire sub URL's. Then another baby crawler in the system crawls each and every collected web pages to get the parsed informations of the each web page. Proposed system extracts the very important features (like title sentences, most repeated words and numerical data etc.) From the user comments or posts by using strong NLP protocols. These feature scores used as as crisp values for fuzzy logic to classify the summary for the recommendation. Then finally by using similarity measures between the 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 get integrated with any social networking sites. This can be done by some parameter settings like

- ✓ Setting up Seed url of the site
- ✓ Setting up Social networking sites API integration for Jason object for https protocol
- ✓ Adding of 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: 2249-9555 Vol. 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 ENGINEERING* VOL:26 NO:7 YEAR 2014

[10] "Personalized Research Paper Recommendation System using Keyword Extraction Based on UserProfile" Kwanghee Hong, , Hocheol Jeon, , Changho Jeon, India.
