INTERNATIONAL JOURNAL FOR ADVANCE RESEARCH IN ENGINEERING AND TECHNOLOGY
WINGS TO YOUR THOUGHTS.....

Recommendation System with Sentimental Analysis using Keyword Search

Susan Thomas[1], Jayalekshmi S.[2]

[1]M.Tech Scholar, LBSITW, Poojappura, Thiruvananthapuram 695012,
greeshmamb91@gmail.com
[2]Associate Professor, LBSITW, Poojappura, Thiruvananthapuram 695012
j.lekshmi.s@gmail.com

***Abstract**: Recommendation or recommender system is a system which provides recommendations to users according to their tastes. They are applied in variety of applications like books, movies, restaurants etc. In the proposed work, a recommendation system for hotel recommendations is considered where a user based collaborative filtering algorithm is used. That is the recommendations are obtained based on the similarity between users with similar tastes and preferences. Here keywords are used to indicate the user's preferences, and then the similarity between the current user and other users with the same keyword in their reviews is considered. The reviews given by the users will have both positive and negative sentiments. So in the proposed work, sentimental analysis on the reviews is done using Naive Bayes, a machine learning technique to distinguish between the positive and negative reviews. Also MongoDB database is used to store the review details. The proposed method aims at providing appropriate recommendations with more accuracy over existing method.*

***Keywords:** Collaborative filtering, MongoDB, Naive Bayes, Recommendation system, Similarity computation*

1. INTRODUCTION

Data mining is one of the most attractive interdisciplinary Data mining is one of the important research area in computer science nowadays. Data mining is the process of extracting useful information from large amount of data. It is of two types, directed data mining and undirected data mining. In directed data mining, a model is build based on the available data and then it will describe rest of the data. Whereas, in undirected data mining some relationship is established among the variables.

In earlier days recommendations was through "word of mouth". As years passed by the amount of information has grown exponentially, which results in data overload. So developed, the recommendation systems. Recommendation systems are systems which provide recommendations to users according to their tastes and it guides the user in a personalized way. Also this system will predict the 'rating' or 'preference' that the user would give to an item.

Even though such systems were developed, still the amount of data is increasing day by day. So there is a need to process and analyze large amount of datasets. So "Big Data" came into prominence. Big Data refers to the datasets whose size is beyond the ability of current technology, method and theory to capture, manage and process the data within a tolerable elapsed time [1]. Nowadays Big Data is a challenge to the IT companies. Recommendation systems can be implemented in Big Data environment.

Sentimental analysis is an important research area in data mining. Sentiment analysis is the study of people's opinions or sentiments towards an entity. It identifies the sentiment expressed in a text and analyzes it.

In this paper, we contribute the following: (1) A user-based collaborative filtering recommendation system (2) Sentiment analysis on the reviews is done using a machine learning technique, Naive Bayes (3) We implement it using MongoDB database.

This paper is organized as follows: Section 2 gives a background. Section 3 describes the user based collaborative filtering algorithm and the sentiment analysis using Naive Bayes. Section 4 gives the analysis of the proposed work from the existing method. Section 5 gives the conclusion and future.

2. BACKGROUND

Recommendation systems have been developed in the mid-1990s [2]. There are a lot of recommendation systems nowadays. Recommendation system is a system that provides recommendations to users.

4.1 Recommendation System

Recommendation system can be classified into content based approach, collaborative filtering approach and hybrid approach. Content based recommendation systems will recommend items based on the description of the items and profile of the user. Collaborative filtering recommendation system will recommend items based on the similarity between the users who have rated the same item before. Hybrid is a combination of content based and collaborative filtering approaches [3].

4.2 Collaborative Filtering(CF)

Collaborative filtering (CF) methods are of two types: item based and user based collaborative filtering. Item based CF, recommend items based on the similarity between the items rated by the same user in the past. User based CF, recommend items based on the similarity between the users who have rated the same items [4]-[5].

4.3 Sentiment Analysis

Sentiment Classification techniques can be divided into machine learning approach, lexicon based approach and hybrid approach. Machine learning approach is based on machine learning algorithms. Lexicon based approach is based on sentiment lexicon and hybrid approach combines both approaches [4].

4.4 MongoDB

INTERNATIONAL JOURNAL FOR ADVANCE RESEARCH IN ENGINEERING AND TECHNOLOGY
WINGS TO YOUR THOUGHTS.....

MongoDB is a document oriented database. It provides high performance, high availability and automatic scaling. A mongoDB instance may have zero or more databases. A database may have zero or more collections. A collection may have zero or more documents. A document may have zero or more fields. It is used to replace the traditional RDBMS for handling large amount of data.
In the proposed method, a user based collaborative filtering algorithm is used for similarity computation and sentiment analysis on reviews is done using Naive Bayes, a machine learning technique and also MongoDB is used for handling unstructured data.

3. PROPOSED METHODOLOGY

In the proposed work, a hotel recommendation system is developed. That is, such a system will recommend the most appropriate hotel to the user in a location according to the user preferences. Here keywords are used to indicate the user preferences and also the candidate services that is, the services provided by the hotels.
In this method two data structures are used, candidate service set and domain thesaurus [1]. Candidate service set is the set of services that a hotel provides. Table 1 shows the candidate service set of a hotel recommendation system.

Table 1: Candidate service set of hotel recommendation system

NO	KEYWORD	NO	KEYWORD
1.	Family	9.	Shopping
2.	Environment	10.	Fitness
3.	Value	11.	Location
4.	Transportation	12.	View
5.	Cleanliness	13.	Quite
6.	Food	14.	Internet
7.	Room	15.	Bar
8.	Service	16.	Beach

Domain thesaurus is the reference work of the candidate service set which contains terms with similar meanings. The domain thesaurus is build manually. Table 2 shows the domain thesaurus of a hotel recommendation system.

Table 2: Domain Thesaurus of hotel recommendation system

Family	Friend, Friends, Family, Daughter, Mother, Father, Son, Child, Wife, Kid
Environment	Modern, Environment, Comfortable, Classic
Value	Price, Cheap, Value, Worth, Money, Expensive, Pay
Transportation	Subway, Stop, Transportation, Bus, Cab, Taxi, Airport, Train, Railway
Cleanliness	Clean, Dirty, Grubby, Neat
Food	Eat, Food, Dishes, Dinner, Breakfast, Delicious, Meal, Restaurant, Lunch
Room	Bed, Room, Bathroom
Service	Service, Waiter, Staff, Server
Shopping	Mall, Shopping, Store, Market
Fitness	Pool, Fitness, Gym, Swimming, Spa
Location	Location
View	View, Scene, Nature, Natural
Quite	Quite, Calm
Internet	Internet, Wifi, Wi-Fi
Bar	Bar
Beach	Beach

3.1 Preprocessing Stage

The first step is the preprocessing stage. In this stage stop words and HTML tags are removed from the reviews. Then the words are converted to the root forms of the word. That is words such as computing, computation and computes have the root form as compute. This process of getting the root form of the words is called stemming. Porter Stemmer algorithm is used for stemming [6]. After stemming the next step is the keyword extraction.

3.2 Keyword Extraction

In this step, the keyword set of the previous user is obtained. That is each review will be converted to the keyword set according to the domain thesaurus and candidate service set. If the review contains a word, which corresponds to a keyword in domain thesaurus, then that keyword is extracted and added to the preference set of the previous user.

3.3 Similarity Computation

After keyword extraction process, the next step is similarity computation. That is to find the neighbors of the active users that have similar tastes to the previous users. A user based collaborative filtering technique is used in this approach. This is done by processing on the reviews of previous users. For example, if a user has rated an item, then the similarity between all other users that have rated the same item is computed. For similarity computation there are many methods. They are Jaccard similarity computation and cosine similarity computation.
Similarity computation between the preferences of active user and previous users is an important part of this algorithm. For similarity computation, when weightage of the keywords are not given, then Jaccard similarity computation is used. Jaccard coefficient is measurement of asymmetric information on binary (and non binary) variables and it is useful when negative values give no information [1].

INTERNATIONAL JOURNAL FOR ADVANCE RESEARCH IN ENGINEERING AND TECHNOLOGY
WINGS TO YOUR THOUGHTS.....

$$sim_{jaccard}(apk, ppk_j) = \frac{|apk \cap ppk_j|}{|apk \cup ppk_j|} \quad (1)$$

where apk is the active user preferences and ppk_j is the previous user's preferences.

When weightage of the keywords are given as input, then Cosine similarity computation is used. Here weightage refers to the importance of the keyword in the reviews. Here in this approach, the preference keyword sets of the active and previous users are converted to their weight vectors [1]. The weight of the keywords given by the user is the active user's preference weights. The previous user's preference weights are computed by the term frequency-inverse document frequency approach.

Term frequency (TF) is the frequency of a particular word in the document [1].

$$TF = \frac{N_{pk_i}}{\sum_g N_{pk_i}} \quad (2)$$

where N_{pk_i} is the number of occurrences of the keyword pk_i in all the keyword sets of the reviews that are given by the same user u, g is the number of keywords in the preference keyword set of the user u.

Inverse document frequency (IDF) is the number of documents divided by the number of documents that contain the word [1].

$$IDF = log \frac{|R'|}{|r: pk_i \in r|} \quad (3)$$

where $|R'|$ is the total number of reviews given by the user u, and $|r: pk_i \in r|$ is the number of reviews where keyword pk_i appears.

After computing the term frequency and inverse document frequency, their product will give the weightage for the keywords contained in the previous user's preferences. Then cosine similarity computation is done using the active and previous user's weights.

$$sim_{cosine}(apk, ppk) = \frac{\sum_{i=1}^{n} \overrightarrow{W}_{ap,i} \times \overrightarrow{W}_{pp,i}}{\sqrt{\sum_{i=1}^{n} \overrightarrow{W^2}_{ap,i}} \sqrt{\sum_{i=1}^{n} \overrightarrow{W^2}_{pp,i}}} \quad (4)$$

where $\overrightarrow{W}_{ap}$ and $\overrightarrow{W}_{pp}$ are the preference weight vectors of active and previous user respectively [1].

3.4 Naive Bayes

Supervised techniques have two sets of data, train data and test data [7]. After the classifier is trained with the training data, the next is feature selection. Here each review is considered as document. When the keywords are given, similarity computation
between users is done. After similarity computation sentimental analysis on the reviews are done. Then the sentence having the keyword from the reviews is taken. Then sentiment analysis on that word is taken using Naive Bayes approach.

Naive Bayes is simple and an effective machine learning algorithm. This is a probabilistic classification algorithm. This is based on Bayes rule. Naive Bayes perform sentimental analysis based on conditional probabilities. This model will not consider the position of the words in the document [8]. The probability that a document d belongs to a class c is:

$$P(c|d) = \frac{P(c)\, P(d|c)}{P(d)} \quad (5)$$

P(d) is a constant if the size of the dataset is known. Therefore P(d) is not considered in calculation for maximum a posteriori. Maximum a posteriori is the most likely class.

$$C_{MAP} = argmax_{c \in C}\, P(d|c)P(c) \quad (6)$$

P(d|c) is the likelihood probability and P(c) is the prior probability. Each document d can be represented as vector of features. That is d = x_1, x_2,…. x_n.

Given the class c, Naive Bayes assumes that each word, wk in the document occurs independently in the document [8].

$$P(c|d)\ \alpha\ P(c)\ \prod_{k=1}^{n_d} [P(w_k|c)]^{t_k} \quad (7)$$

where n_d is the number of unique words in document d and t_k is the frequency of each word w_k . When applying Naive Bayes classifier (NBC), we can estimate P(c) and P(w_k jc) as [8]:

$$\widehat{P_C} = \frac{N_c}{N} \quad (8)$$

$$P(\widehat{w_k}|c) = \frac{count\,(w_k|c)+1}{\sum_{w \in V} count(c) + |V|} \quad (9)$$

where N is the total number of documents, N_c is the number of documents in class c and $|\boldsymbol{V}|$ is the vocabulary.

After computing the sentimental weightage value, it is added to the personalized rating. Personalized rating is computed as [1]:

$$PR = \overline{r} + k \sum_{ppk_j \in \hat{R}} sim\,(apk, ppk_j) \times (r_j - \bar{r}) \quad (10)$$

where,

$\bar{\boldsymbol{r}}$ - average rating of the services

sim(apk,ppk_j) - similarity between active user preference and previous user preferences

k - normalized value which is the reciprocal of sum of similarity between apk and ppk_j

r_j - rating of a particular service j

4. FLOWCHART FOR THE PROPOSED SYSTEM

Figure 1 shows the flowchart for the proposed system. The active user will give the preferences and its weightage. Once it is given the system checks whether the weightage is zero

INTERNATIONAL JOURNAL FOR ADVANCE RESEARCH IN ENGINEERING AND TECHNOLOGY
WINGS TO YOUR THOUGHTS.....

or not. If it is zero, then jaccard similarity computation is done. Otherwise it performs cosine similarity computation.

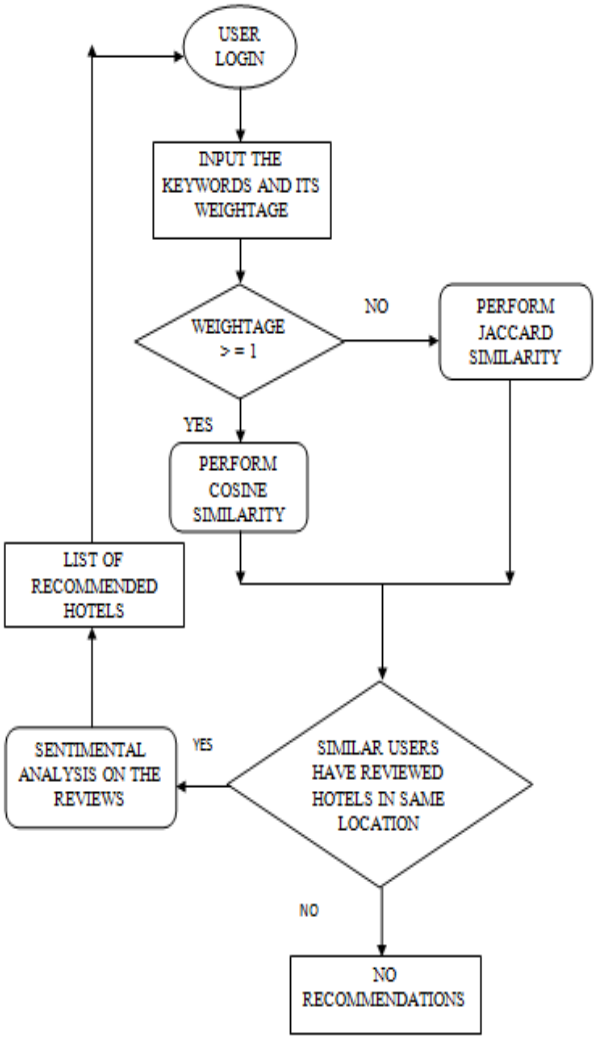

Figure 1: Proposed System

After similarity computation, it checks whether the similar users have reviewed hotels in the same location as given by the active user. If so, then sentimental analysis on those reviews is done. Then the list of hotels is recommended to the active user. If similar users have not reviewed hotels in same location given by the active user, then no recommendations are obtained.

5. MONGODB

Relational databases were used widely in many applications and it provides good performance when they handle limited amount of data. To handle large amount of data such as internet, social media etc it was di_cult. So inorder to overcome this problem , came the NoSQL databases. NoSQL was coined by Carlo Strozzi in 1998 and refers to non relational databases.

The advantage of a NoSQL database is that, it can handle unstructured data such as documents, email, multimedia and social media efficiently, which is not possible in relational databases [9]. Nonrelational databases do not have the RDBMS principles (Relational Data Base Management System). They do not store data in tables, schema is not fixed and have very simple data model. Instead, they use keys and data can be identified based on the keys assigned.

MongoDB is a document oriented database. It provides high performance, high availability and automatic scaling. A mongoDB instance may have zero or more databases. A database may have zero or more collections. A collection may have zero or more documents. A document may have zero or more fields [9].

- Documents in the same collection dont even need to have the same fields
- Documents are the records in RDBMS
- Documents can embed other documents
- Documents are addressed in the database via a unique key

Table 3: Comparison of MongoDB and MySQL [9]

MongoDB	**MySQL**
Collection	Table
Document	Row
Field	Column
Embedded Documents, Linking	Join

MongoDB is used nowadays for storing large amount of data. Even though, it had advantages and disadvantages [9]. The advantages are sharding, speed and flexibility. The disadvantages are no join, memory usage high and concurrency issues.

6. EXPERIMENTAL EVALUATION AND RESULTS

The dataset used is the Tripadvisor dataset. This dataset consists of reviews and ratings given by the users for the hotels. In this system, MongoDB is used to store the review details. The dataset consists of hotels in 8 locations and has total 2998 reviews with 1135 users.

Mean Absolute Error (MAE) and Normalized Mean Absolute Error (NMAE) is used for evaluation of the results in this work. MAE [10] is a statistical accuracy metric often used in CF methods to measure the prediction quality. It is defined as the average absolute deviation between a predicted rating and the real rating.

$$MAE = \frac{\sum_{i=1}^{M} |q_i - p_i|}{M} \quad (11)$$

where q_i and p_i are the real rating and the corresponding predicted rating respectively, and M is the number of the pairs of real ratings and predicted ratings $< q_i, p_i >$. The Normalized Mean Absolute Error (NMAE) metric [10] is also used to measure the prediction accuracy, which is defined as:

$$NMAE = \frac{MAE}{\frac{\sum_{i=1}^{M} q_i}{M}} \quad (12)$$

The lower the MAE or NMAE presents the more accurate predictions.

INTERNATIONAL JOURNAL FOR ADVANCE RESEARCH IN ENGINEERING AND TECHNOLOGY
WINGS TO YOUR THOUGHTS.....

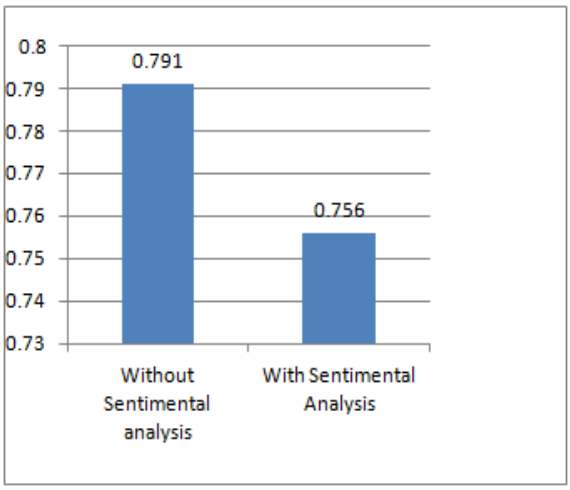

Figure 2: MAE values using cosine similarity computation

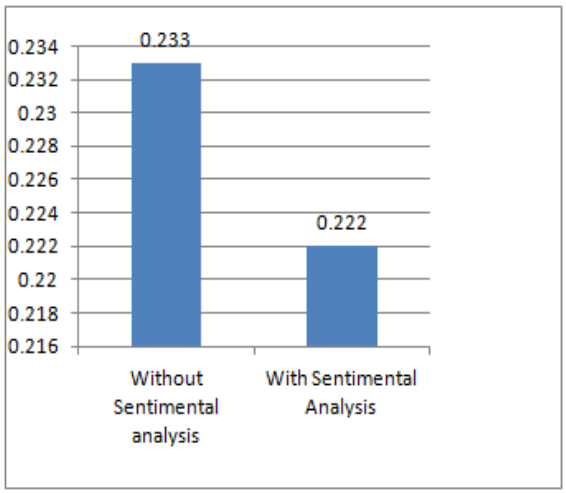

Figure 3: NMAE values using cosine similarity computation

Figure 2 shows that the MAE values for Cosine similarity computation with sentimental analysis is 4.42 percent lower than without sentimental analysis and figure 3 shows that the NMAE values for Cosine similarity computation with sentimental analysis is 4.31 percent lower than without sentimental analysis. This means that MAE and NMAE values with sentimental analysis is lower, so much more accurate the predictions.

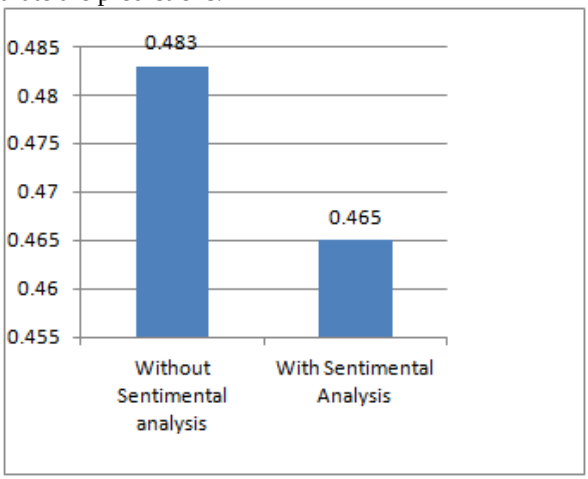

Figure 4: MAE values using jaccard similarity computation

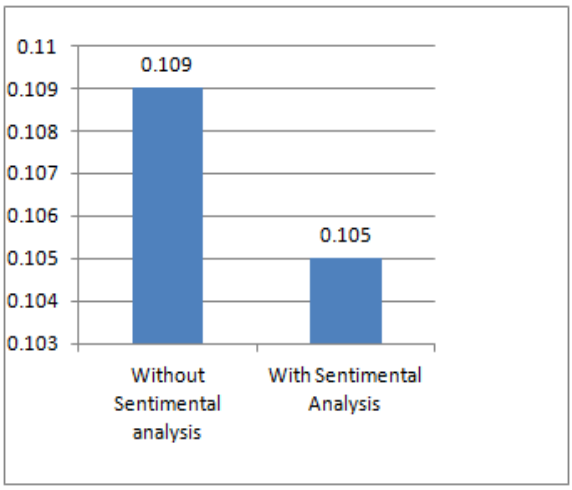

Figure 5: NMAE values using jaccard similarity computation

Figure 4 shows that the MAE values for Jaccard similarity computation with sentimental analysis is 3.73 percent lower than without sentimental analysis and figure 5 shows that the NMAE values for Jaccard similarity computation with sentimental analysis is 3.67 percent lower than without sentimental analysis. This means that MAE and NMAE values with sentimental analysis is lower, so much more accurate the predictions.

From this it is clear that Cosine similarity computation shows better performance than Jaccard similarity computation. And also the proposed work has lower MAE and NMAE values for both the similarity computation. Therefore the proposed work provides an improvement in the accuracy.

7. CONCLUSION AND FURTURE WORK

Recommendation or recommender system is a system which provides recommendations to users according to their tastes. They are applied in variety of applications like books, movies, restaurants etc. In this work a hotel recommendation system is implemented. Here similarity between users are computed using similarity computation. And also sentimental analysis on the reviews are done using Naive Bayes, a machine learning technique.

For similarity computation cosine and jaccard similarity computation is used. And the results shows that cosine similarity computation shows better performance that jaccard similarity computations. Also in the proposed system, MongoDB is used as a big data solution. It is used to handle unstructured data that is to handle the reviews given by the users. It replaces the traditional relational database. MongoDB is a document oriented database. This is more efficient than MySQL.

In future work, other supervised or unsupervised machine learning algorithms can be used for sentimental analysis. Also instead of creating the domain thesaurus manually, in future it can be created dynamically so more accurate results can be obtained.

INTERNATIONAL JOURNAL FOR ADVANCE RESEARCH IN ENGINEERING AND TECHNOLOGY

WINGS TO YOUR THOUGHTS.....

ACKNOWLEDGMENT

I feel unequivocal gratification in acknowledging the service and co-operation rendered by Mrs. Jayalekshmi S., Associate Professor, Department of Computer Science and Engineering, for giving valuable suggestions and guidance throughout the completion of my work.

I also acknowledge with grateful thanks the authors of the references and other literature referred to in this paper.

References

[1] **Shunmei Meng, Wanchun Dou, Xuyun ZhangPlata and Jinjun Chen, "KASR: Keyword Aware Service Recommendation Method on MapReduce for Big Data Applications", *IEEE Transactions on parallel and distributed systems*, pp.1-11, 2014.**

[2] **Paul Resnik and Hal R. Varian, "Recommnder systems", communications of the ACM, vol.40, No.3,pp.56-58,March 1997.**

[3] **G. Adomavicius, and A. Tuzhilin, "Toward the Next Generation of Recommender Systems: A Survey of the State of-the-Art and Possible Extensions", *IEEE Transactions on Knowledge and Data Engineering*, Vol.17, No.6 pp. 734-749, 2005.**

[4] **Zhi-Dan Zhao and Ming-Sheng Shang, "User-based Collaborative-Filtering Recommendation Algorithms on Hadoop", *In the third International Workshop on Knowledge Discovery and Data Mining*, pp. 478-481, 2010.**

[5] **Gui-Rong Xue, Chenxi Lin, Qiang Yang and WenSi Xi, "Scalable Collaborative Filtering Using Cluster-based smoothing", *In. proceedings of 2005 ACM*, pp.144-121.**

[6] **Biju Issac and Wendy J. Jap, "Implementing Spam Detection using Bayesian and Porter Stemmer Keyword Stripping Approaches", *TEN-CON 2009-2009 IEEE Region 10 Conference*, pp. 1-5, 2009.**

[7] **Walaa Medhat, Ahmed Hassan b and Hoda Korashy, "Sentiment analysis algorithms and applications: A survey", *Production and hosting by Elsevier B.V. on behalf of Ain Shams University.*, pp. 1093-1113, Elsevier, April 2014.**

[8] **Bingwei Liu, Erik Blaseh, Yu Chen, Dan Shen and Genshe Chen, "Scalable Sentiment Classification for Big Data Analysis Using Naive Bayes Classifier", *IEEE Intl Conf. on Big Data*, October 2013.**

[9] **Cornelia Gyorodi, Robert Gyorodi, George Pecherle, Andrada olah, "A Comparative Study: MongoDB vs. MySQL", IEEE Conference, June 2015.**

[10] **Kleanthi Lakiotaki, Nikolaos F. Matsatsinis and Alexis Tsoukias, "Multi-Criteria User Modeling in Recommender Systems", IEEE Intelligent System, Vol.26, No.2 pp. 64-76, 2011.**