ISSN: 2229-6956 (ONLINE)
DOI: 10.21917/ijsc.2016.0171

PRS: PERSONNEL RECOMMENDATION SYSTEM FOR HUGE DATA ANALYSIS USING PORTER STEMMER

T.N. Chiranjeevi[1] and R.H. Vishwanath[2]
[1]*Department of Computer Science and Engineering, Sambhram Institute of Technology, India*
E-mail: chiranjeevitn@gmail.com, vishwa.gld@gmail.com

Abstract
Personal recommendation system is one which gives better and preferential recommendation to the users to satisfy their personalized requirements such as practical applications like Webpage Preferences, Sport Videos preferences, Stock selection based on price, TV preferences, Hotel preferences, books, Mobile phones, CDs and various other products now use recommender systems. The existing Pearson Correlation Coefficient (PCC) and item-based algorithm using PCC, are called as UPCC and IPCC respectively. These systems are mainly based on only the rating services and does not consider the user personal preferences, they simply just give the result based on the ratings. As the size of data increases it will give the recommendations based on the top rated services and it will miss out most of user preferences. These are main drawbacks in the existing system which will give same results to the users based on some evaluations and rankings or rating service, they will neglect the user preferences and necessities. To address this problem we propose a new approach called, Personnel Recommendation System (PRS) for huge data analysis using Porter Stemmer to solve the above challenges.
In the proposed system it provides a personalized service recommendation list to the users and recommends the most useful services to the users which will increase the accuracy and efficiency in searching better services. Particularly, a set of suggestions or keywords are provided to indicate user preferences and we used Collaborative Filtering and Porter Stemmer algorithm which gives a suitable recommendations to the users.
In real, the broad experiments are conducted on the huge database which is available in real world, and outcome shows that our proposed personal recommender method extensively improves the precision and efficiency of service recommender system over the KASR method. In our approach mainly consider the user preferences so it will not miss out the any of the data, based on the ranking system and gives better preferential recommendations.

Keywords:
Recommender System, Preferences, Similarity, Data Reduction, Prediction

1. INTRODUCTION

Data mining is the process of retrieving or extracting the patterns, information, and rules etc., which are not in the database. Data mining is also an interdisciplinary subfield of software engineering. The main objective of the data mining is to enhance the business by making use of extracted information into the Business system. Data mining also includes database and database administration perspectives, information pre-handling, model and induction contemplations, interestingness measurements, unpredictability contemplations, post processing of found structures, representation and web overhauling.

Data mining is widely used in Financial Data Analysis to manage the account information and also the budgetary information of the industry. It is also used to handle credit and debit based information and day today transaction process.

Data Mining has its incredible application in Retail Industry as it gathers substantial measure of information from the user mainly on deals, sales, client buying history, merchandise transportation, utilization and administrations.

Today the telecom business is a standout amongst the most developing commercial ventures giving different administrations, for example, fax, pager, phone, web delivery person, pictures, email, web information transmission etc., to handle all these kind of information we need data mining.

Recently, we have seen a huge development in the field of Bioinformatics, for example, genomics, proteomics and biomedical examination. Organic information mining is a vital piece of Bioinformatics.

The applications examined above tend to handle generally small and homogeneous information sets for which the statistical strategies are suitable. Tremendous measure of information have been gathered from different domains, for example, geosciences, astronomical field and so on. A huge amount of data is being updated in different fields, for example, atmosphere or climate related information and in different engineering fields etc.

In the last few years, huge amount of data is being generated from various sources; such generated mountain of data has to be properly analyzed. The analysis of such mountain of data is known as Big Data analytics. Big Data refers to database or the dataset whose size is much more farer than the ability of current technology, and the methods used and it is a theory to confine and process the data within a stipulated time. Now a days, big data management became a challenging task for IT companies. The key to solve such a radical thing is by changing the hardware and specifying more convenient software solutions. Big Data also provides new prospect and vital challenges to information technological industry and academic institutions [7] [8].

Big data possesses the following characteristics:

Volume: Here volume is referred to the amount of data which is generated, and that is significant in the context of Big Data. The potential and value of the data which is under review and also whether this treated as Big Data or not is dependent on the size of the data. 'Big Data' the name itself has a term associated with the size and so is its characteristic.

Variety: Successive characteristic is the variety of the Data. It is critical fact that analysts must know that to which category the Big Data should fit in. Categorising the data will help the people who are strictly examining the data which is relevant with it, to use it fruitfully and therefore preserving the status of Big Data.

Velocity: Velocity is the term with the frame of reference to the quickness of data or at what rate the data can be processed, rendered generated in order to meet the threats which will lie further in the pathway of development and growth.

Variability: This is characteristic which may become an issue for those who will strictly go through the data. This implies to the conflicts that can be displayed by the Big Data at some times, therefore obstructing the process handling to manage data impressively.

Veracity: The reputation of the captured data can differ greatly. Analysing the accuracy will be depending on the variety of data source.

Complexity: Managing the data can become very tedious process, particularly when the volume of data is very large and from many sources. Data has to be linked or chained to one another so as to make sure that it is being conveyed in proper order.

Likewise most of the big data applications, the huge information slant additionally make substantial effects on administration recommender frameworks. With the increment in the quantity of substitute administrations, productive prescribing administrations that benefactors favour has turn into an imperative exploration concern. Administration recommender frameworks have been broadcasted as vital devices to help supporters manage administrations weight and present suitable proposal to them. For example usage of Digital versatile disk or compact disk, books, web access and various other stuffs now use recommender systems. But in the last few years, there been much research done both in data innovative industry and scholarly organizations on growing new routines for administration recommender frameworks.

1.1 MOTIVATION

The main motivation behind this work is that it is required to provide most accurate key information to clients to take up wise decision. This can be achieved by developing an efficient service recommender system. The service recommender systems are most use full tools for providing suitable suggestions to the patrons. Over the time, the quantity of clients, services and usage of online information has developed in a brisk manner, creating the big data analysis problem for service recommender systems. As a result, the old versions of service recommender frameworks frequently bear the versatility and wastefulness issues when taking care of or investigating enormous information. In this way, the vast majority of existing administration recommender frameworks gives the same appraisals and assessments of administration recommendations to distinctive supporters without considering their assorted nature and overlook their inclinations, and thus neglects to give patrons individual needs.

1.2 METHODOLOGY

In this we primarily used an indexing method called Hash-map technique for faster search and to select the appropriate keywords from the reviews. Also the indexing method is used to eliminate the articles like the, is, a, an etc. from the paragraph.

We also used the Porter Stemmer method mainly for the stemming of keywords from the uploaded reviews. Also this will match the keywords with the user preferences and predict the appropriate recommendations.

1.3 CONTRIBUTION

In this paper, we contributed an approach called Personnel Recommendation System (PRS) for huge data analysis using Porter Stemmer for addressing major issues of both the rating services and the user personal preferences. It provides a domain thesaurus of the system for the active users which helps the users to provide their personal preferences. Based on these preferences and service recommendation list it recommends the most useful services to the users which will increase the accuracy and efficiency in searching better services. Particularly, we used Collaborative Filtering and Porter Stemmer algorithm which gives a suitable recommendations to the users.

1.4 ORGANIZATION

The rest of the paper is organized as follows, Section 2 discusses briefly the Literature on recommendation, similarity, data reduction and prediction. Section 3 presents the background work, Section 4 contains Problem definition, Section 5 describes the System architecture, section 6 presents the Mathematical model and Algorithm, Section 7 addresses the Experimental Results for proposed method and existing KASR technique. Concluding remarks are summarized in the Conclusion.

2. RELATED WORK

Literature review plays an important part in the field of research and development. For an application development, it is necessary to undergo every aspect of it. Previous studies are the source from which research ideas are drawn and developed into concepts and theories. We performed survey mainly on different recommendations systems, Similarity search engines, Data reduction and Prediction system.

2.1 RECOMMENDED SYSTEMS

Recommender frameworks grew as an autonomous re-seek range in the mid-1990s when suggestion issues began concentrating on appraising models.

G. Linden et al. [1] proposed a recommender framework which can be portrayed as structure that makes individualized recommendations as yield or has the effect of dealing with the customer in an altered way to intriguing or valuable administrations in an extensive space of conceivable alternatives. Current proposal strategies for the most part can be characterized into three primary classes: traditional collaborative filtering, cluster model and search based approaches. Substance based methodologies prescribe administrations like those the client favoured previously. Collaborative Filtering (CF) methodologies recommend the client by considering the suggestions from the previous clients who used these recommendations.

G. Linden et al. also suggested an item-to-item collaborative filtering, which scales to huge information sets and delivers top notch suggestions continuously. Rather than matching the user to similar customers, item-to-item collaborative filtering matches each of the clients purchases and rate the similar items, then combines those similar items into a recommendation list. But this method is not suitable in the retail industry to extensively apply recommendation algorithms for focused marketing, both online and offline.

M. Bjelica et al. [2], proposed a recommender system for the broadcast scenario, where uplink connection to the network center is not available. They introduced special prominence on user

modelling algorithm that would be able to efficiently learn the user's interests and give the appropriate recommendations.

M. Bjelica et al. also suggested two major approaches which are content-based and collaborative recommendations. Content-based systems recommend prescribe those things that look like the ones the client preferred before, while collaborative systems recommend the items in which the alternate clients with comparative tastes preferred previously. The main limitation is it does not have more advanced clustering techniques which have the capacity to demonstrate different area of user interests.

M. Alduan et al. [3], proposed a recommender system or game recordings which will be connected in an Olympic Games situation utilizing access from the Internet or broadcast services. The proposed recommender system considers the various variables that can make a client like a specific sports video of the Olympic Games. They also proposed a new recommendation method that is straightforward to the client, who just needs to consume recordings as he/she would do in any video dispersion stage. The framework considers how the preferences of clients change after some time.

2.2 SIMILARITY

The algorithm suggested by M. Bjelica et al. [2] generates recommendations based on a few customers who are most similar to the patrons. It can measure the similarity of two patrons, in different methods; a common method is to measure the cosine of the angle between the two vectors. The algorithm can select recommendations from the similar patrons or various methods as well, a common technique is to categorize each item according to how many similar patrons purchased it.

G. Linden et al. [1] also proposed a technique to find customers who are similar to the patrons, cluster models divide the client base into many parts and treat the job as a classification problem. Here the goal is to allocate the patron to the segment containing the most similar clients. It then uses the purchases and ratings of the clients in the segment to produce recommendations. Once the algorithm generates the parts, it computes the patrons similarity to vectors that summarize each parts, then chooses the segment with the appropriate similarity and categorize the patrons accordingly. They suggested that the cluster models which group many clients together in a part, match a patron to a segment, and then consider all clients in the segment who are similar to the patrons for the purpose of making recommendations. The main limitation in this is that the similar clients found by the cluster models are not the most suitable similar patrons, the recommendations they produce are less relevant.

G. Linden et al. [1] recommended a technique to determine the most-similar match for a given product, they suggested an algorithm that builds a similar-products table by finding products that clients tend to purchase together.

According to G. Linden et al., P. Castells et al. and Y. Zhu et al. [1], [9], [10] it's possible to calculate the similarity between the products in various ways, but a common method is to use the cosine measure, in which each vector corresponds to a product rather than a client, and the vector's dimensions correspond to clients who have purchased that product.

Kalpa Gunaratna1 et al. [11] proposed a new approach for document recovery that makes use of predication mined from the documents. Here author measure the similarity between predications to compute the similarity between documents. Author computed the similarity between two predications as the average pair wise similarity of topic, predicate, and items duo. They make use of various level relationships between ideas, as well as levels of relationships, to measure concept similarity. The main limitation in this is they used SemRep which uses a template-based predication mining and hence it can miss extraction of some predications from articles.

2.3 DATA REDUCTION

G. Linden et al. [1] suggested that, it is possible to partly address the level of issues by reducing the data size. He also suggested that we can also reduce the data size by randomly sampling the clients or discarding clients with few purchases, and reduce data size by discarding very popular or unpopular items. It is also possible to reduce the number of products observed by a small, constant factor by partitioning the item space based on product category or subject categorization.

J. Dean et al. [12] proposed a model which uses the reduce function that accepts an intermediate key from the client and a set of values for that key. It merges together these values to form a possibly smaller set of values.

G. Linden et al. [1] suggested that dimensionality decrease systems connected to the product space tend to have the same impact by wiping out low-recurrence products. The main limitation in this is the dimensionality diminishment connected to the client space successfully bunches comparable clients into groups, such grouping can lower recommendation quality.

S. Ghemawat et al. [13] proposed that the chunk size is one of the key design parameters. A large chunk size offers several important advantages. First, it reduces users need to interact with the main server because it reads and writes on the same chunk which require only one original request to the main server for chunk location information. The reduction is particularly important for the workloads because application mostly read and writes large files in an orderly manner.

2.4 PREDICTION

Z. Zheng et al. [14], proposed a QoS ranking prediction framework for cloud services by considering benefit of the past service usage knowledge of other clients. Their proposed framework requires no extra invocations of cloud services when making QoS ranking prediction. Two personalized QoS ranking prediction approaches are proposed to predict the QoS rankings directly. The simplest approach of personalized cloud service QoS ranking is to calculate all the candidate services at the client side and categorize the services based on the observed QoS values. He also proposed a personalized ranking prediction framework, named CloudRank, to predict the QoS ranking cloud service without requiring other real world facility invocations from the intended clients. Their approach takes benefits of the previous usage knowledge of other patrons for making personalized ranking prediction for the active patrons.

They also proposed a rating-oriented collaborative filtering approach. First it predict the missing QoS values before making QoS ranking. The main limitation is that the present approach only rank dissimilar QoS properties separately.

Mejdl S. Safran et al. [15] propose a new knowledge-based approach to improve significance prediction in focused Web crawlers. For this study, they chose Naïve Bayesian as the main prediction model. A focused crawler specifically searches and downloads only the web pages that are appropriate to a particular subject. In order to search web pages that are appropriate to a particular topic, a focused crawler recognize and track links that guide to appropriate pages, and disregard links that direct to inappropriate pages. The main limitation in this is it does not support other categorization models like Support vector machine and neural networks.

Jian Hu et al. [16] study the dilemma of predicting web user's gender and age based on their browsing activities, in which the webpage outlook information is considered as a concealed variable to broadcast demographic information between various patrons. Firstly, based on patrons summary and their browsing record, the patrons age and gender knowledge is broadcasted to the browsed pages and then an administered regression model is qualified to predict a Webpage's gender and age affinity. Secondly, within Bayesian structure, an internet patron's age and gender are predicted based on the age and gender affinity of the Webpages that he/she has searched. Here the main limitation is it only predict information based on gender and age it will miss out other attributes like occupation, location etc.

Jiannan Wang et al. [17] study the difficulty of automatic URL completion and prediction using fuzzy type-ahead search which useful for the patrons who mistype the address in the URL. Using fuzzy search is very significant when the patron has incomplete information about URLs. As the patrons types keywords their technique predicts appropriate URLs that include words similar to the enquired keywords. The main limitations of this is it cannot predict the URL that have not been used by the patrons.

3. BACKGROUND

Shunmei Meng [6] proposed a technique called "KASR: A Keyword-Aware Service Recommendation Method on Map Reduce for Big Data Applications" for providing the recommendation for the users which addresses the problem of rating or ranking services based on the keywords and recommendations. The authors implemented this model using Map Reduce Technique which uses vector weight method to analyse and predict the recommendations. This entire process consumes more time to provide correct recommendations to the users.

To overcome above limitation in this paper, we propose a frame work called "PRS: Personnel Recommendation System for Huge Data Analysis using Porter Stemmer". In this we used the different techniques called Hash Map and Porter Stemmer. Hash Map directly eliminates the articles present in the huge database and Porter Stemmer predicts the accurate recommendation for the users.

Table.1. Symbols and Notations

L_{ou}	Reviews of old users
O_w	Keyword from the old user reviews
L_{pu}	Preferential list of present user
P_w	Keyword from the preferential list
$Sim(L_{ou}, L_{pu})$	Similarity between L_{ou} and L_{pu}
P_R	Personalized recommendation list
R	Recommended keyword

4. PROBLEM STATEMENT

Given a huge database $D = \{D_1, D_2, ..., D_n\}$, in this database each D_i contains number of different existing users preferential use full keywords. Thus, we consider keyword candidate list which is a set of preferential or most useful keywords about user preferences and multi criteria of the candidate services, of the old or previous users can be denoted as,

$$L_{ou} = \{ow_1, ow_2, \ldots, ow_n\} \quad (1)$$

where, ow_i is the number of the useful words or keywords of old or previous users in the keyword-candidate list.

Similarly, we consider keyword candidate list which is a set of preferential or most useful keywords about user preferences and multi criteria of the candidate services, of the present or active users can be denoted as,

$$L_{pu} = \{pw_1, pw_2, \ldots, pw_n\} \quad (2)$$

where, pw_i is the number of the useful words or keywords of the present or active users in the keyword-candidate list.

By considering these two keyword list such as L_{ou} and L_{pu} the objective is to suggest personalized service recommendation list to the present or active user and recommend the most appropriate services to the current user. Also, suggest a custom-made service recommendation catalogue and to recommend the most suitable service(s) to the patrons.

5. SYSTEM ARCHITECTURE

The system architecture is depicted in Fig.1 and has the following components,

i) Input Unit
ii) Pre-processing Unit
iii) Prediction Engine
iv) Output Unit

Input Unit: The input unit consists of (i) List of Preferences of Old Users and (ii) List of Preferences of Present User.

The List of Preferences of Old Users consists of keyword candidate list which is a set of preferential or most useful keywords about user preferences and multi criteria of the candidate services, of the old or previous users which can be denoted as, $L_{ou} = \{ow_1, ow_2, ..., ow_n\}$.

The List of Preferences of Present User consists of set of preferential or most useful keywords about user preferences and multi criteria of the candidate services, of the present or active user which is denoted as, $L_{pu} = \{pw_1, pw_2, ..., pw_n\}$.

Pre-processing Unit: The list of preferences from both Present and old users are given as input to the Pre-processing unit. The main task of this unit is to remove the stop words and HTML tags from the User preferential Reviews collections to get the good quality keywords. Then the Porter Stemmer method is applied to

remove the inflexional endings and commoner morphological words in English.

Another main task of this unit is to perform keyword extraction. In this step, all reviews are transformed into domain thesaurus and keyword-candidate list. If the same word found in the both review list and domain thesaurus, then that keyword is extracted and placed into the preference keyword set of the user.

In this, we also eliminate the reviews which are not related to present user preferences by the concept called intersection in the set theory. There by reducing the size of the Reviews list for further operations.

Prediction Engine: Prediction Engine performs mainly three tasks. Namely (i) Similarity Computation, (ii) Computing Personalized Ratings and (iii) Generating Recommendations.

The Similarity Computation is performed by applying the Jaccard coefficient measure to find out the reviews of the Present and old users have the same tastes of similarity of their preferences.

A personalized rating is calculated by using indexing method in which it uses the frequency of keywords in the reviews mainly to calculate the repetitions of keywords.

Finally, based on these frequencies all the chosen keywords are added into the Array list A_L. By considering this array list A_L appropriate personalized recommendation is generated.

We can also compute the Inverse Document Frequency (IDF) by dividing the number of all reviews by the number of reviews containing the keyword *ow*.

Output Unit: The output unit finally contains the list of personalized service recommendations with top ratings.

6. MATHEMATICAL MODEL AND ALGORITHM

In our method, two data structures, "keyword-candidate list" and "specialized domain thesaurus", are introduced to help obtain user preferences.

The keyword-candidate list is a set of keywords about users preferences and multi criteria of the candidate services, which can be denoted as $K = \{k_1, k_2, ..., k_n\}$, where n is the number of the keywords in the keyword-candidate list.

An area thesaurus is a reference work of the pivotal word hopeful rundown that rundowns words assembled together as indicated by the likeness of decisive word significance, including related and differentiating words and antonyms.

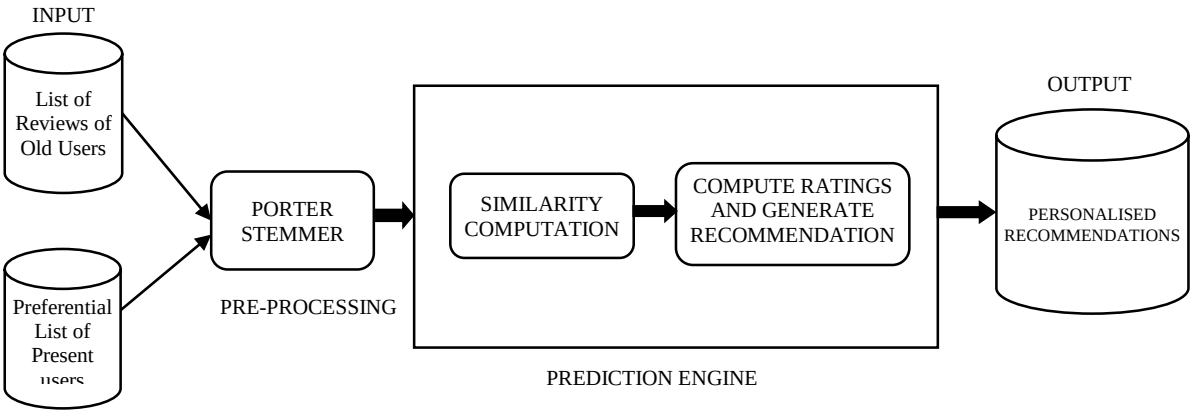

Fig.1. System Architecture

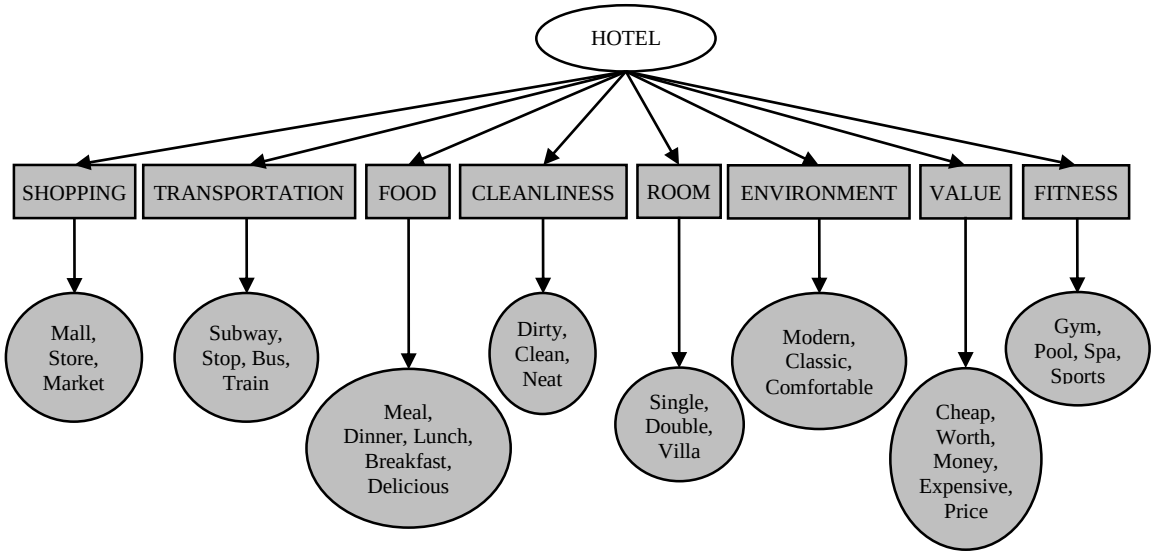

Fig.2. A Domain Thesaurus of Hotel

A specimen of a thesaurus of hotel reservation structure is exhibited in Fig.2. As demonstrated in Fig.2, the words in the red rectangle are the crucial words in the looking at catchphrase cheerful summary, and the words in the ovals are the related articulations of the fundamental words. Habitually, space thesauruses are updated reliably to ensure the advantageousness of the words.

For example, if a review of an old client for a hotel has the word "bus", which is corresponding to the keyword "Transportation" in the domain thesaurus, then the keyword "Transportation" should be contained in the preference keyword set of the previous user.

The List of Preferences of Old Users which can be denoted as $L_{ou} = \{ow_1, ow_2, ..., ow_n\}$ and the List of Preferences of Present User which is denoted as, $L_{pu} = \{pw_1, pw_2, ..., pw_n\}$. Then the Porter Stemmer method is applied to remove the inflexional endings and commoner morphological words in English. Then extracting the keywords using Hash map technique to obtain the key word list K_L,

$$KL = \{kw_1, kw_2, \ldots, kw_n\} \quad (3)$$

Then Similarity Computation is performed by applying the Jaccard coefficient as follows,

$$Sim\left(L_{ou}, L_{pu}\right) = \frac{\left|L_{ou} \cap L_{pu}\right|}{\left|L_{ou} \cup L_{pu}\right|} \quad (4)$$

where, L_{ou} and L_{pu} are the list of old and present users keywords. The frequency of keyword in the keyword set is given by,

$$F = \frac{K_r}{\sum_i K_u}. \quad (5)$$

Here F is the frequency of keywords, ow represents the number of occurrences of keywords in the reviews commented by the users. Pw is the preferential keyword set.

We can also compute the inverse document frequency (IDF) by dividing the number of all reviews by the number of reviews containing the keyword ow.

$$IDF = \log \frac{|R_v|}{|r : k \in r|} \quad (6)$$

where, $|L_{ou}|$ is the total number of the reviews commented by user, and $|r{:}ow \in r|$ is the number of reviews where keyword k appears.

Finally, we generate recommendation list that contains the list of personalized service recommendations with top n ratings. such as,

$$P_R = \{R_1, R_2, R_3, \ldots, R_K\}. \quad (7)$$

The detail algorithm steps are discussed in Table.2.

Table.2. Algorithm: PRS

Input: Old user review list L_{ou} and present user preferential list L_{pu}.
Output: Personalized service recommendation list $P_R = \{R_1, R_2, R_3, ..., R_n\}$
Begin
Phase I (Pre-processing)
//Reading preferences of both Present and Old users and Extracting the keywords.
1. *Initialize* $L_{ou} = 0$, $L_{pu} = 0$, $K_L = 0$;
2. *Read* L_{ou}
3. *Read* L_{pu}
4. for each review list $L_{ou} = \{ow_1, ow_2, ..., ow_n\}$
5. extract the keywords $K_L = \{kw_1, kw_2, ..., kw_n\}$ then
6. for the preference list of present user
7. $L_{pu} = \{pw_1, pw_2, ..., pw_n\}$
8. if $L_{ou} \cap L_{pu} \neq \theta$ then
9. insert L_{ou} into R_K into buffer
10. end if
11. end for
12. end for
13. return L_{ou}

Phase II (Prediction engine)
//Similarity Computation, Computing Personalized ratings and generating recommendations
14. for each keyword set $L_{ou} \in R_K$
15. $Sim\left(L_{ou}, L_{pu}\right) = \frac{\left|L_{ou} \cap L_{pu}\right|}{\left|L_{ou} \cup L_{pu}\right|}$
16. if $Sim(L_{ou}, L_{pu}) < \Delta$ then
17. remove L_{ou} from R_K
18. else insert the word into the array list A_L
19. end if
20. end for
21. if *Sim*!= null then
22. add similar words to the array list A_L
23. end if
24. Provide the recommendations according to the personalized ratings generated for preferences L_{pu}
25. return the services with the Top K personalized recommendation list $P_R = \{R_1, R_2, R_3, ..., R_K\}$

End

First we read the preferences of both present and old users from the preferential list of the present user and from the review statement of the old users respectively which can be shown from step 1 to step 4 from the Table.2.

Then extract the keywords from the old user reviews to compare it with the preferences of the present users which is shown in step 5. After extracting the keywords porter stemmer will stem the keywords to match it with the preferences of present users which is described from step 6 to step 9.

After stemming we calculate similarity computation between stemmed words and the present user preferences and this is handled in the steps from 14 to step 18.

After similarity computation we add the keywords to the array list to give the appropriate recommendations which is shown in the step 21 to 25.

7. EXPERIMENTAL EVALUATION AND RESULT

In this section, experiments are conducted on real time hotel review datasets which are collected from various sources in the

web. To evaluate the performance of our proposed method PRS, we compare with other recommendation method such as: Keyword Aware Service Recommendation method on MapReduce (KASR). Here, we used three metrics to evaluate the accuracy: (i) Mean Absolute Error (MAE) [18], (ii) Mean Average Precision (MAP) [19] and (iii) Discounted Cumulative Gain (DCG) [20].

The Mean Absolute Error is given by,

$$MAE = \frac{1}{N}\sum_{i=1}^{N}|R' - R| \quad (8)$$

where, R' is the predicted value of the recommended system. R is current value of the recommender system and N is the number of review words used for prediction.

The Mean Average Precision is given by,

$$MAP = \frac{\sum_{i=1}^{n} P \times R}{N} \quad (9)$$

where, P is the predicted value, R is the rank and N is the number of reviews.

The Discounted Cumulative Gain is given by,

$$DCG = \Sigma 2^{I} - 1/\log(1+P) \quad (10)$$

where, i is the predicted value of the recommendation and P is the rank of the predicted value.

7.1 ACCURACY EVALUATION

7.1.1 Comparison of KASR and PRS in MAE:

MAE is a statistical accuracy metric often used in Collaborative Filtering (CF) method to measure the prediction quality. The lower the MAE presents the more accurate predictions.

Table.3. Performance of KASR and PRS with respect to MAE

MAE (KASR)	MAE (PRS)
0.1042	0.0774
0.0926	0.0595
0.9060	0.0586
0.0866	0.0554
0.0800	0.054
0.0752	0.0531
0.0745	0.0525
0.0721	0.0522
0.0718	0.0482
0.0698	0.0443

The Fig.3 and Table.3 shows the MAE computed values for both KASR and PRS. It is observed that the MAE values of PRS is much lower than KASR (e.g., the MAE of PRS is (((0.1042-0.0774)/0.1042) = 25.71%) Thus our method PRS provide more accuracy in prediction than existing KASR method.

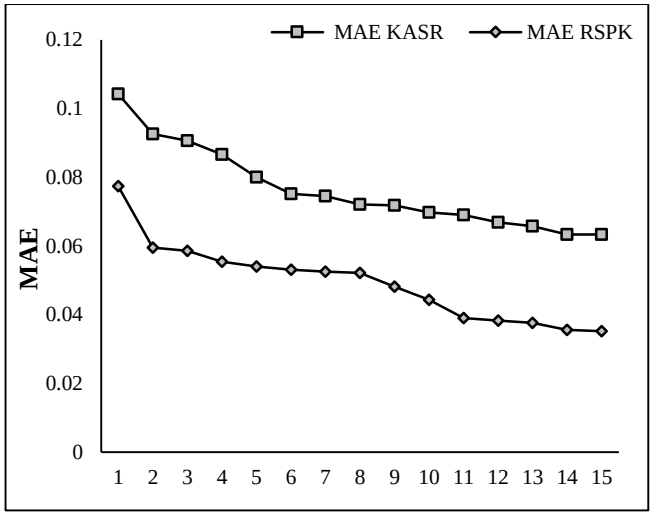

Fig.3. Comparison of KASR and PRS in MAE

Table.4. Performance of KASR and PRS with respect to DCG

DCG (KASR)	DCG (PRS)
7.93	10.82
3.82	6.01
2.96	4.69
2.42	3.85
2.12	3.18
1.9	2.77
1.78	2.54
1.64	2.32
1.48	2.22
1.29	2.04

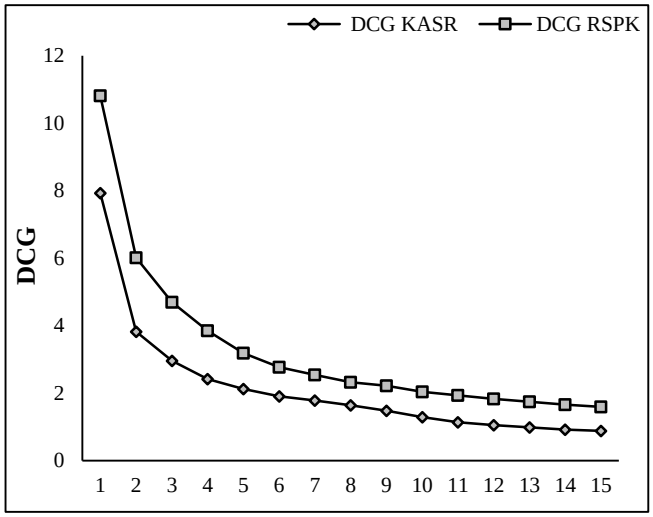

Fig.4. Comparison of KASR and PRS in DCG

7.1.2 Comparison of KASR and PRS in MAP and DCG:

To evaluate the quality of Top-K service recommendation list, MAP and DCG are used as performance evaluation metrics.

Higher the MAP or DCG values then it presents higher quality of predicted service recommendation list. The related values obtained are shown in the Table.4 and Fig.4 for the evaluation metric DCG. And the related values obtained for evaluation metric are shown in the Fig.5 and Table.5.

Table.5. Performance of KASR and PRS with respect to MAP

MAP (KASR)	MAP (PRS)
0.2572	0.3316
0.3575	0.4516
0.3532	0.4174
0.3602	0.4118
0.3250	0.4059
0.2938	0.3993
0.2953	0.4041
0.2760	0.3937
0.3286	0.4401
0.3653	0.4806

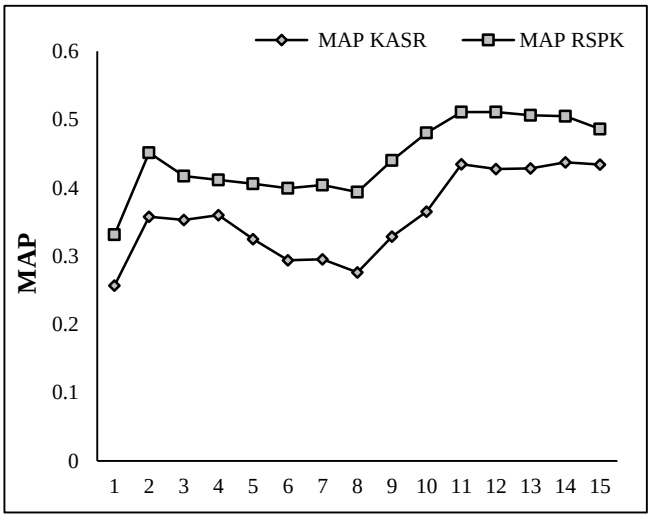

Fig.5. Comparison of KASR and PRS in MAP

Table.6. Average of KASR and PRS with respect to MAE MAP and DCG

	AVG MAE	AVG DCG	AVG MAP
KASR	0.0764	2.153	0.3583
PRS	0.0494	3.279	0.4437

8. CONCLUSION AND FUTURE WORK

The proposed mechanism PRS works in two phase process. In the first phase we perform pre-processing task by capturing the preferences of both old and present users from that preferences we mainly extract the use full key words. In the second phase, the prediction of personalized recommendation services are provided by similarity computation using Jaccard coefficient method and also by computing personalized ratings by computing the frequency of occurrences of keywords present in the reviews collected from the old users.

Further, our experimental results show that the average MAE is 0.0764%, average DCG is 2.153% and average MAP is 0.3583% in the existing KASR method, whereas in the proposed method, the average MAE is 0.0494%, average DCG is 3.279% and average MAP is 0.4437% which is observed in the Table.6. Thus, the proposed method is 3.5% more efficient with respect to MAE, 3.4% more efficient for DCG and 1.9% more efficient for MAP compared to existing KASR method.

In our future work, we can deal with the circumstance where term appears in particular groupings of a space thesaurus from association and how to recognize the positive and negative slants of the customers from their comments to determine more precise recommendation.

REFERENCES

[1] G. Linden, B. Smith and J. York, "Amazon.com Recommendations: Item-to-Item Collaborative Filtering", *IEEE Internet Computing*, Vol. 7, No. 1, pp. 76-80, 2003.

[2] Milian Bjelica, "Towards TV Recommender System Experiments with User Modeling", *IEEE Transactions on Consumer Electronics*, Vol. 56, No. 3, pp. 1763-1769, 2010.

[3] Faustino Sanchez, Maria Alduan, Federico Alvarez, Jose Manuel Menendez and Orlando Baez, "Recommender System for Sport Videos Based on User Audiovisual Consumption", *IEEE Transactions on Multimedia*, Vol. 14, No. 6, pp. 1546-1557, 2012.

[4] 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 Knowledge and Data Engineering*, Vol. 17, No. 6, pp. 734-749, 2005.

[5] B Badrul Sarwar, George Karypis, Joseph Konstan and John Riedl, "Item-Based Collaborative Filtering Recommendation Algorithms", *Proceedings of 10th International Conference on World Wide Web*, pp. 285-295, 2001.

[6] Shunmei Meng, Wanchun Dou, Xuyun Zhang and Jinjun Chen, "KASR: A Keyword-Aware Service Recommendation Method on MapReduce for Big Data Applications", *IEEE Transactions on Parallel and Distributed Systems*, Vol. 25, No. 12, pp. 3221-3231, 2014.

[7] Fay Chang, et al., "Bigtable: A Distributed Storage System for Structured Data", *ACM Transactions on Computer Systems*, Vol. 26, No. 2, pp. 1-14, 2008.

[8] Wanchun Dou, Xuyun Zhang, Jianxun Liu and Jinjun Chen, "HireSome-II: Towards Privacy-Aware Cross-Cloud Service Composition for Big Data Applications", *IEEE Transactions on Parallel and Distributed Systems*, Vol. 26, No. 2, pp. 455-466, 2013.

[9] Pablo Castells, Miriam Fernandez and David Vallet, "An Adaptation of the Vector-Space Model for Ontology-Based Information Retrieval", *IEEE Transactions on Knowledge and Data Engineering*, Vol. 19, No. 2, pp. 261-272, 2007.

[10] Yingwu Zhu and Yiming Hu, "Enhancing Search Performance on Gnutella-Like P2P Systems", *IEEE*

Transactions on Parallel and Distributed Systems, Vol. 17, No. 12, pp. 1482-1495, 2006.

[11] Kalpa Gunaratna, "Document Retrieval using Predication Similarity", Wright State University, 2016.

[12] Jeffrey Dean and Sanjay Ghemawat, "MapReduce: Simplified Data Processing on Large Clusters", *Communications of the ACM*, Vol. 51, No. 1, pp. 107-113, 2005.

[13] Sanjay Ghemawat, Howard Gobioff and Shun-Tak Leung, "The Google File System", *Proceedings of 19th ACM Symposium on Operating Systems Principles*, pp. 29-43, 2003.

[14] Zibin Zheng, Xinmiao Wu, Yilei Zhang, Michael R. Lyu and Jianmin Wang, "QoS Ranking Prediction for Cloud Services", *IEEE Transactions on Parallel and Distributed Systems*, Vol. 24, No. 6, pp. 1213-1222, 2013.

[15] Mejdl S. Safran, Abdullah Althagafi and Dunren Che "Improving Relevance Prediction for Focused Web Crawlers", *Proceedings of IEEE/ACIS 11th International Conference on Computer and Information Science*, pp. 161-166, 2012.

[16] Jian Hu, Hua-Jun Zeng, Hua Li, Cheng Niu and Zheng Chen "Demographic Prediction Based on User's Browsing Behavior", *Proceedings of the 16th International Conference on World Wide Web*, pp. 151-160, 2007.

[17] Jiannan Wang, Guoliang Li, Jianhua Feng, Chen Li "Automatic URL Completion and Prediction Using Fuzzy Type-Ahead Search", *Proceedings of the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 634-635, 2009..

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

[19] Yi Chen. Pan and Lin Shan Lee, "Performance Analysis for Lattice-Based Speech Indexing Approaches Using Words and Subword Units", *IEEE Transactions on Audio, Speech, and Language Processing*, Vol. 18, No. 6, pp. 1562-1574, 2010.

[20] Guosheng Kang, Jianxun Liu, Mingdong Tang, Xiaoqing Liu, Buqing Cao and Yu Xu, "AWSR: Active Web Service Recommendation Based on Usage History", *Proceedings of IEEE 19th International Conference on Web Services*, pp. 186-193, 2012.