AN EFFECTIVE APPROACH OF CONTENT BASED RECOMMENDATION METHOD ON BIG DATA APPLICATIONS

1. **Ilakkiya[1]** 2. **A.Prema[2]**

1. M.Phil., Scholar, Department of Computer Science, Raja Doraisingam Arts College, Sivaganga.

2. Assistant Professor, Department of Computer Science, Raja Doraisingam Arts College,

ABSTRACT

Now days, online searching process increases and people searches new information in the search process. Most of the search engine gives additional supporting information. Recommender system involves in this process and implements as service. Service recommender system gives additional information to the user but if information grows then these process become a critical one. The proposed work analyses issues occurring when service recommender system implements in large data sets. This work proposes a keyword-Aware services Recommender method, to split the services to the users and mainly focused keywords from the user preferences. Collaborative filtering algorithm generates keyword recommenders from the previous user preferences. To implement effective results in big data environment, this method is implemented by the concept of MapReduce parallel processing on Hadoop. Experimental results are shown the effective results on real-world datasets and reduce the processing time from large dataset.

KEYWORD: Big Data, Map Reduce, Hadoop.

I. INTRODUCTION

Big data is a blanket term for the non-traditional strategies and technologies needed to gather, organize, process, and gather insights from large datasets. While the problem of working with data that exceeds the computing power or storage of a single computer is not new, the pervasiveness, scale, and value of this type of computing has greatly expanded in recent years. Big data is a "large dataset" that means a dataset too large to reasonably process or store with traditional tooling or on a single computer. This means that the common scale of big datasets is constantly shifting and may vary significantly from organization to organization. These datasets can be orders of magnitude larger than traditional datasets, which demands more thought at each stage of the processing and storage life cycle. Another way in which big data differs significantly from other data systems is the speed that information moves through the system. Data is frequently flowing into the system from multiple sources and is often expected to be processed in real time to gain insights and update the current understanding of the system.

Big data problems are often unique because of the wide range of both the sources being processed and their relative quality. Data can be ingested from internal systems like application and server logs, from social media feeds and other external APIs, from physical device sensors, and from other providers. Big data seeks to handle potentially useful data regardless of where it's coming from by consolidating all information into a single system.

Big data also brings new opportunities and critical challenges to industry and academia. Similar to most big data applications, the big data development also positions heavy impacts on service recommender systems. With the number of alternative services, effectively recommending services that users preferred have become an important research issue. Service recommender systems provided valuable tools to help users deal with services overload and appropriate recommendations. Recommender systems apply techniques and methodologies from another neighboring areas -such as Human computer interaction (HCI)or Information Retrieval(IR). However, most of these systems bear in their core an algorithm that can be understand as a particular instance of a data mining (DM) technique. Recommender System makes use of different sources of information for providing users with predictions and recommendations of items. They try to balance factors like accuracy, novelty, dispersity and stability in th

e recommendations. Collaborative Filtering (CF) methods play an important role in the recommendation, although they are often used along with other filtering techniques like content-based, knowledge-based or social ones. CF is based on the way in which humans have made decisions throughout history: besides on our own experiences, we also base our decisions on the experiences and knowledge that reach each of us from a relatively large group of acquaintances.

II. RELATED WORK

The administration suggestion framework gives extra data to the client. As of now the measure of information has expanded quickly that yields enormous information issue. In conventional administration system confronts two issues when handling vast measure of information. They are proficiency of information and versatility issue. Additionally the vast majority of the customary prescribed frameworks concentrate on the positioning and appraisals of administration with distinctive administrations and diverse client. Existing framework does not consider the client inclination, just spotlights on the thing of the administration.

Shunmei Meng et.al. (2013) proposed the administration proposal framework utilized the Collaborative Filter calculation to create the surmised suggestion to the dynamic client. That enhances the versatility and effectiveness of the information when transforming extensive measure of information. Works are executed on Hadoop utilizing Map Reduce structure with constant information set. Finally results are compared to the existing system result of the recommendation system. And to provides accurate data and scalability of large amount of data. The recommendation system predicts the future user taste based on the active user preference. Rapid growth of data the user cannot pick out the required data in internet.

Reema Sikka et.al. (2012) discussed about the filtering approaches and different types of recommendation system. To improve the accuracy of the data and also used the collaborative filter to predict the user taste to generate the recommendation to the active user. And the item based approaches to identify the relation between the past and present user preference, also implemented the user based approach to predict the user taste. Finally, to compare the different algorithms such as Random prediction, Frequent Sequence, Collaborative Filter, and Content based filter. The collaborative filter algorithm was used. The collaborative filter algorithm is to produce efficient data and improve the scalability problem. The recommendation system e-learning is based on the user data and evaluation of result. Currently, Resources are available bulk of learning material in online or offline. The peoples are selecting the correct required material in internet. Also study about various recommendation techniques to explain the four filters, they are Demographic filter, Content based filter, Collaborative filter and Hybrids filter.

Rubina Parveen et.al. (2012) discussed the drawbacks and advantages of the entire filter. The viewers choices on the product advertisement are important part of the market. This part generates the recommendation to the viewers. Atisha Sachan et.al.(2012) is proposed the two data collection methods. They are implicit and explicit data collection. Also the filter can be classified into four types, they are 1.Demographic filter, 2.Content based filter, 3. Collaborative filter, 4. Hybrids filter. And discuss about the four filters but mainly focused on the more powerful and effective collaborative filter. The major challenges of this filter are 1.Cold start problem, 2.Data Sparsity, 3.Scalability, 4.Accuracy of data problem. A proposal framework in an e-learning connection is endeavour's to brilliantly prescribe to the dynamic client (learner) in light of the activity of the past client (learner). Likewise this proposal framework is in light of the online movement, for example, perusing posted message on web. These proposal frameworks have been attempted in an e-business to buy the quality products.

Dhoha Almazro et.al (2013) locations the utilization of web mining to construct the online exercises taking into account client (learner) access History to enhance course material route and enhance the transforming time. Additionally elearning contrasted with the conventional arrangement of eye to eye style showing and learning. It gives a larger number of profits than the eye to eye learning. Here the content based Filter is utilized. The content based Filter to utilize the calculation of vector space model for similitude processing to foresee the comparative thing taking into account the client evaluated thing that enhance the Scalability of information. The data about the item is expanding with exponential rate in e-trade industry .

III. METHODOLOGIES IN RECOMMENDER SYSTEM

The past client's surveys are put away into the Data set. The Data set was downloaded from UCI archive. The measure of the Data set is 257MB. The Data set was put away in the content records. Principle characteristics of the Data set is Hotel Name, Year, Previous client Reviews. The principle capacity of pre-processing is to expel the commotion from the past client audits and obliged catchphrases are removed from the information set utilizing watchman stemmer calculation. Porter stemmer calculation: The Porter stemmer calculation is utilized to expel the postfix word from the past clients survey. What's more the regular morphological additionally first class that term called stem. The archive speaks to the term or vector structure. Here the past client audit essential words are grouped utilizing watchman stemmer calculation.

Similarity Checking: The past client surveys are removed from Dataset and put away the database. The characteristics are Hotel name, Date, magic words, check. The dynamic client gives their obliged Keywords with rating. The Previous User Keyword (PUK) are contrasted with the Active User Keywords (AUK). In the event that the decisive words are precisely same then the catchphrases are put away in the framework utilizing Pearson connection coefficient calculation this system is called precise comparability Otherwise the strategy is said to be rough closeness.

We accumulated the current client's audits and put away it into the information set in fig 1. The Data set was downloaded from UCI storehouse. It is named as Hotel Reservation System. The qualities of this information set are Hotel name, Date of the Review and Users audit. The pre-processing step is utilized to expel the commotion from information set. In the wake of evacuating the clamour the information is put away into database. These courses of action are under the disconnected from the net. At that point the Active client gives their favoured thing and appraised of the thing.

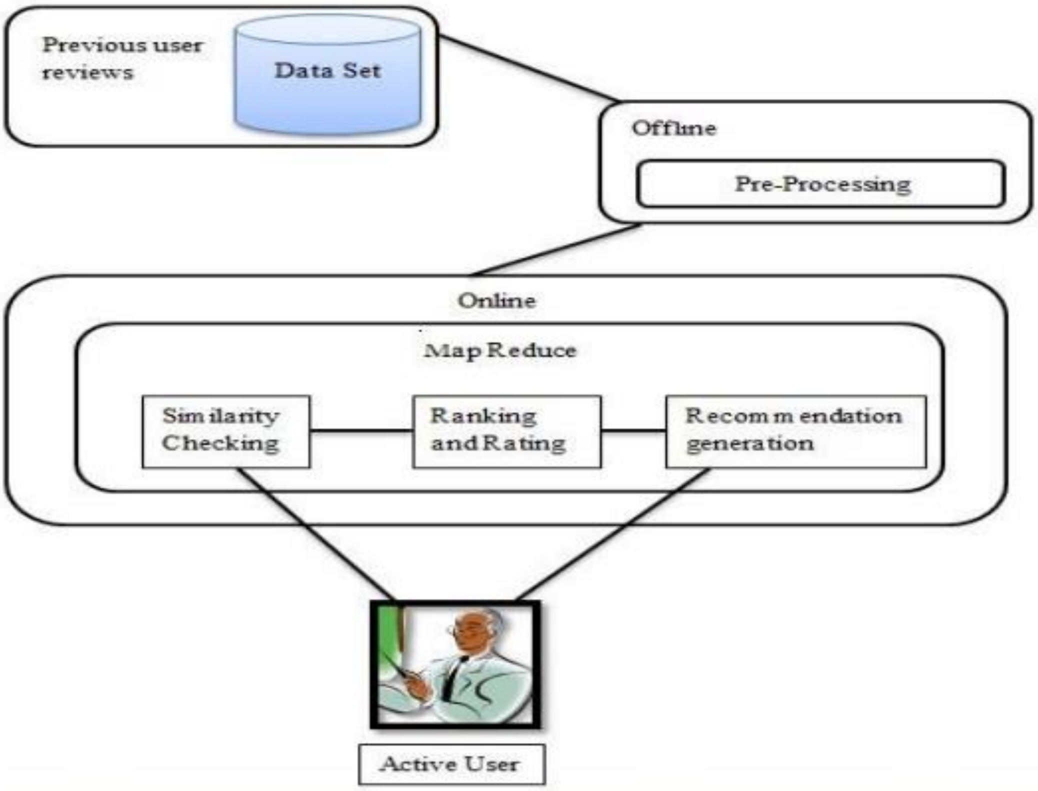


Fig1.Structure of processing work

That favored thing was contrasted with the past client inclination. On the off chance that the Active client favoured thing is like the past client thing then the obliged comparable thing is appraised [20] and positioned. At long last we produce the Recommendation to the Active client. Online environment closeness checking, Ranking and Rating, Recommendation Generation [8] are performed utilizing Map decrease idea.

The process for generating an RS recommendation is based on a combination of the following considerations:

- The type of data available in its database (e.g., ratings, user registration information, features and content for items that can be ranked, social relationships among users and location-aware information).
- The filtering algorithm used (e.g., demographic, content-based, collaborative, social-based, context-aware and hybrid). The model chosen (e.g., based on direct use of data: ‘‘memory based,’’ or a model generated using such data: ‘‘model-based’’).
- The employed techniques are also considered: probabilistic approaches, Bayesian networks, nearest neighbors algorithm; bio-inspired algorithms such as neural networks and genetic algorithms; fuzzy models, singular value decomposition techniques to reduce sparsity levels, etc.
- Sparsity level of the database and the desired scalability.
- Performance of the system (time and memory consuming).
- The objective sought is considered (e.g., predictions and top N recommendations) as well as The desired quality of the results (e.g., novelty, coverage and precision). Resear

3.1 Content-Based Filtering:

Content-based filtering makes recommendations based on user choices made in the past (e.g. in a web-based e-commerce RS, if the user purchased some fiction films in the past, the RS will probably recommend a recent fiction film that he has not yet purchased on this website). Content-based filtering also generates recommendations using the content from objects intended for recommendation; therefore, certain content can be analyzed, like text, images and sound. From this analysis, a similarity can be established between objects as the basis for recommending items similar to items that a user has bought, visited, heard, viewed and ranked positively.

3.2 Demographic filtering:

Demographic filtering is justified on the principle that individuals with certain common personal attributes (sex, age, country, etc.) will also have common preferences.

3.3 Collaborative Filtering:

Collaborative Filtering allows users to give ratings about a set of elements (e.g. videos, songs, films, etc. in a CF based website) in such a way that when enough information is stored on the system, we can make recommendations to each user based on information provided by those users we consider to have the most in common with them. CF is an interesting open research field. As noted earlier, user ratings can also be implicitly acquired (e.g., number of times a song is heard, information consulted and access to a resource).

3.4 Hybrid Filtering:

Hybrid filtering commonly uses a combination of CF with demographic filtering or CF with content-based filtering to exploit merits of each one of these techniques. Hybrid filtering is usually based on bioinspired or probabilistic methods such as genetic algorithms, fuzzy genetic, neural networks, Bayesian networks , clustering and latent features.

IV. METRICS IN RECOMMENDER SYSTEMS

Research in the RS field requires quality measures and evaluation metrics to know the quality of the techniques, methods, and algorithms for predictions and recommendations. Evaluation metrics and evaluation frameworks facilitate comparisons of several solutions for the same problem and selection from different promising lines of research that generate better results.

Evaluation metrics] can be classified as (a) prediction metrics: such as the accuracy ones: Mean Absolute Error (MAE), Root of Mean Square Error (RMSE), Normalized Mean Average Error (NMAE); and the coverage (b) set recommendation metrics: such as Precision, Recall and Receiver Operating Characteristic (ROC)] (c) rank recommendation metrics: such as the half-life and the discounted cumulative gain] and (d) diversity metrics: such as the diversity and the novelty of the recommended items. The validation process is performed by employing the most common cross validation techniques (random sub-sampling and k-fold cross validation) ; for cold-start situations, due to the limited number of users (or items) votes involved, the usual method chosen to carry out the experiments is leave-one out cross validation.

n order to measure the accuracy of the results of an RS, it is usual to use the calculation of some of the most common prediction error metrics, amongst which the Mean Absolute Error (MAE) and its related metrics: mean squared error, root mean squared error, and normalized mean absolute error stand out.

The novelty evaluation measure indicates the degree of difference between the items recommended to and known by the user. The diversity quality measure indicates the degree of differentiation among recommended items.

The stability in the predictions and recommendations influences on the users' trust towards the RS. A RS is stable if the predictions it provides do not change strongly over a short period of time.

4.1 Results and Analysis

The past client's surveys are put away into the Data set. The Data set was downloaded from UCI archive. The measure of the Data set is 257MB. The Data set was put away in the content records. Principle characteristics of the Data set is Hotel Name, Year, Previous client Reviews. The principle capacity of pre-processing is to expel the commotion from the past client audits and obliged catchphrases are removed from the information set utilizing watchman stemmer calculation.

The past client surveys are removed from Dataset and put away the database. The characteristics are Hotel name, Date, magic words, check. The dynamic client gives their obliged Keywords with rating. The Previous User Keyword (PUK) are contrasted with the Active User Keywords (AUK). In the event that the decisive words are precisely same then the catchphrases are put away in the framework utilizing Pearson connection coefficient calculation this system is called precise comparability Otherwise the strategy is said to be rough closeness.

Set based algorithm is utilized to figure out the likeness between the dynamic client magic words and the past client decisive word.

$$\text{Sim (AUK, PUK)} = \text{AUK} \cap \text{PUK} / \text{AUK U PUK}$$

4.2 Exact Similarity (Collaborative Filter):

The obliged decisive words of the dynamic client and past client watchwords are changed into the type of n dimensional weight vector. Furthermore Similarity and weight was computed utilizing Pearson Correlation coefficient calculation.

V. MAP REDUCE:

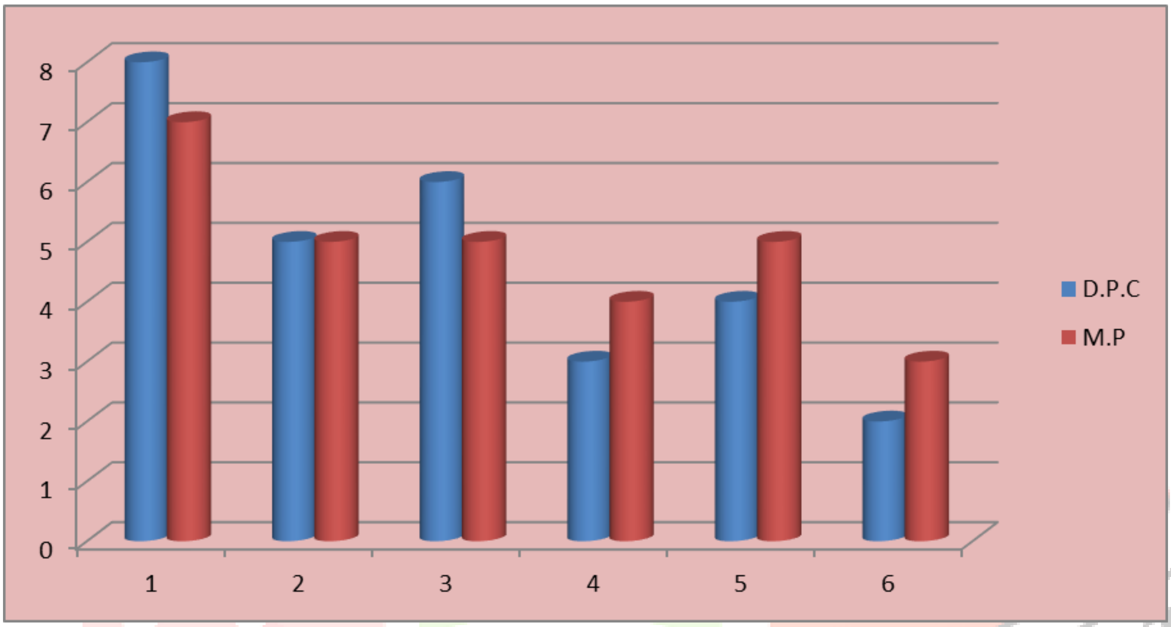
Map Reduce [1] [3] [23] is utilized to execute the information in parallel way. Here the Similarity checking, weight era, rating

5.1 Scalability:

Exploratory results show the productivity of the proposed Pearson Correlation calculation.

Table 1. Demonstrates the examination of cosine based algorithm and the Pearson Correlation coefficient calculation.

NO.	Detection Probability count	Max Value Probability	Processing Time
1.	8	7	0.6232
2.	5	5	0.4861
3.	6	5	0.5391
4.	3	4	0.3172
5.	4	5	0.4126
6.	2	3	0.2831



VI. CONCLUSION

Recommender system involves in this process and implements as service. Service recommender system gives additional information to the user but if information grows then these process become a critical one. The proposed work analyses issues occurring when service recommender system implements in large data sets. This work proposes a keyword-Aware services Recommender method, to split the services to the users and mainly focused keywords from the user preferences. This paper analyses recommender system methodologies and evaluates its metrics. Collaborative recommender system predicts relevant results from user's data and reduces search time. This method has greater flexibility compared to other methods and achieves optimal results.

REFERENCES

[1] Michael J. Pazzani and Daniel Billsus, "Content-based recommendation systems," Springer Berlin Heidelberg. pp. 325-341,2007

[2] Badrul Sarwar, George Karypis, Joseph Konstan, and John Ried, "Item-based Collaborative Filtering Recommendation Algorithms," May 1-5, 2001, Hong Kong.

[3] R. Burke, "Hybrid Recommender Systems: Survey and Experiments," User Modeling and User-Adapted Interaction, vol. 12, no. 4, pp. 331-370, 2002.

[4] L. Sharma and A. Gera, "A Survey of Recommendation System: Research Challenges," International Journal of Engineering Trends and Technology (IJETT), vol. 4, May 2013.

[5] Shunmei Meng, Wanchun Dou, Xuyun Zhang,"KASR: A Keyword-Aware Service Recommendation Method on Mapreduce for Big Data Applications," IEEE Trans. on Parallel and Distributing systems, vol.25, no.12, December 2014.

[6] Y. Chen, A. Cheng, and W. Hsu, "Travel Recommendation by Mining People Attributes and Travel Group Types from Community-Contributed Photos," IEEE Trans. Multimedia, vol. 25, no. 6, pp. 1283-1295, Oct. 2013

[7] M. Alduan, F. Alvarez, J. Menendez, and O. Baez, "Recommender System for Sport Videos Based on User Audiovisual Consumption," IEEE Trans. Multimedia, vol. 14, no. 6, pp. 1546-1557,Dec. 2012.

[8] S. Alonso, F.J. Cabrerizo, F. Chiclana, F. Herrera, E. Herrera-Viedma, Group decision making with incomplete fuzzy linguistic preference relations, International Journal of Intelligent Systems 24 (2009) 201–222.

[9] Ansari, S. Essegaier, R. Kohli, Internet recommendation systems, Journal of Marketing Research 37 (3) (2000) 363–375.

[10] N. Antonopoulus, J. Salter, Cinema screen recommender agent: combining collaborative and content-based filtering, IEEE Intelligent Systems (2006) 35– 41.

[11] P. Antunes, V. Herskovic, S.F. Ochoa, J.A. Pino, Structuring dimensions for collaborative systems evaluation, ACM Computing Surveys 44 (2) (2012). Article 8.

[12] O. Arazy, N. Kumar, B. Shapira, Improving Social Recommender Systems, Journal IT Professional 11 (4) (2009) 31–37.

[13] L. Ardissono, A. Goy, G. Petrone, M. Segnan, P. Torasso, INTRIGUE: Personalized recommendation of tourist attractions for desktop and handset devices, Applied Artificial Intelligence 17 (8-9) (2003) 687–714.

[14] R. Baeza-Yates, B. Ribeiro-Neto, Modern Information Retrieval, AddisonWesley, 1999.

[15] M. Balabanovic, Y. Shoham, Content-based, collaborative recommendation, Communications of the ACM 40 (3) (1997) 66–72.

[16] L. Baltrunas, T. Makcinskas, F. Ricci, Group recommendation with rank aggregation and collaborative filtering, in: Proceedings of the 2010 ACM Conference on Recommender Systems, 2010, pp. 119–126