Scientific Journal Impact Factor (SJIF) : 3.518
International Journal of Modern Trends in Engineering and Research
ISSN (ONLINE) : 2349 - 9745 ISSN (PRINT) : 2393 - 8161

PERSONALIZED INFORMATION RETRIEVAL SYSTEM USING MAP REDUCE AND VECTOR SPACE MODEL

Shivangi Goel[1] And Abhishek Sharma[2]
[1]Mtech Student and [2]Assistant Professor
Department Of Computer Science And Engineering, SRMUniversity, India

Abstract-The big data is the concept of large spectrum of data, which is being created day by day. In recent years handling these data is the biggest challenge. Hadoop is an open source platform which is used effectively to handle the big data applications. The two core concepts of the hadoop are Mapreduce and Hadoop distributed file system (HDFS). HDFS is the storage mechanism and map reduce is the programming language. Results are produced faster than other traditional database operations.We proposed vector space model algorithm and map reduce,this algorithm for as improve the data classification and make it uniform. Then apply modified K-Means clustering on input data which we get from above algorithm and output is stored in clustered form. K means reduce the number of comparison which makes execution faster. Clustered Data act as input for MapReduce. MapReduce apply Mapper,Combiner and Reducer Mechanism over data and eliminate duplicate data from large amount of data set.For test data the divide and conquer approach isapplied on each row of the cluster. Divide andconquer technique is used to match records within a cluster which further improves the efficiency of the algorithm.Web-based recommender that makes suggestions by using text categorization from Search. Recommendation systems are one of these tools. They suggest items of interests (such as books, movies, CDs, news, pictures, etc.) by using statistical and machine learning techniques.

Keyword- Information retrieval, K-means, Modified K-means, Hadoop, Vector space Model, Personalized, Information, Retrieval.

I. Introduction

A MIDST a constantly developing e-market environment, the existence and sustainability of commerce and business is contingent upon maintaining a competitive advantage through effective and aggressive marketing strategies. An abundant amount of information is continuously being created and made accessible through the electronic media. Users, however, lack an adequate tool to help them organize the unwieldy situation. Due to explosive growth of information over the internet in last several decades, information overload is becoming a big challenge. It has also become difficult for users to access relevant information efficiently. Meanwhile, the substantial increase in the number of websites presents a challenging task for webmasters to organize the contents of the websites to cater to the needs of users. Modeling and analyzing web navigation behavior is helpful in understanding what information online user's demand. 'Information filtering' is a rapidly evolving method being used to manage large information flows. The fundamental objective of "information filtering" is to only expose users to information that would be relevant to them. Recommender system facilitates successful e-marketing by focusing on aspects of bettering customer relationship, creating communities of interest and most importantly, building trust. Analyzing and modeling web navigation behavior would greatly enhance understanding of the preferences of on-line users.

We proposed **MapReduce** technique that allows computation to run on a cluster. It uses HDFS for data-proximity.The computation will be distributed and run in parallel on the cluster and each process will access and process data that is available locally on its node. This gives a major performance boost.**VectorSpace Models**were developed to eliminate many of the problems associated with exact, lexical matching techniques. It is difficult for a lexical matching technique to differentiate between two documents that share a given word, but use it differently, without understanding the context in which the word was used. Vector-space models, by placing terms,

documents, and queries in a term-document space and computing similarities between the queries and the terms or documents, allow the results of a query to be ranked according to the similarity measure used. Unlike lexical matching techniques that provide no ranking or a very crude ranking scheme, the vector-space models, by basing their rankings on the Euclidean distance or the angle measure between the query and terms or documents in the space, are able to automatically guide the user to documents that might be more conceptually similar and of greater use than other documents.

Over the last decade, lots of researchers have studied new approaches of personalized information retrieval systems by using Vector Space Model and Modified K-Means, and apply to real world. Especially, applications of MapReduce technique to recommender systems have been effective to offer personalized information to the user through analyzing his/her preference.

II. Related Work

Every company advertises its product on the internet throughWeb advertising service but to advertise according to consumer preferences was difficult. So **Tung-Yen Lai et al [2010]** analyzed this problem and proposed a personalized Web advertisement selection and recommendation system through **a membership based advertisement marketing website**whose advertisement content is determined by consumer preference. As the consumer preference declines with time, **the half-life theory and fuzzy theory is** applied to analyze browsing behaviourof users. Now the target advertisement is filtered to match with the user preferences. Now only those advertisements shown that satisfy users need. Therefore, this increases the advantages of target marketing and market of advertisement is strengthened.

III. Description of the algorithm

A). Vector Space Model:

The vector space model procedure can be divided in to three stages.

i. Document Indexing

It is obvious that many of the words in a document do not describe the content, words like the, is. By using automatic document indexing those non significant words (function words) are removed from the document vector, so the document will only be represented by content bearing words. This indexing can be based on term frequency, where terms that have both high and low frequency within a document are considered to be function words. In practice, term frequency has been difficult to implement in automatic indexing. Instead the use of a stop list which holds common words to remove high frequency words (stop words), which makes the indexing method language dependent. In general, 40-50% of the total number of words in a document is removed with the help of a stop list.

ii. Term Weighting

Term weighting has been explained by controlling the exhaustivity and specificity of the search, where the exhaustivity is related to recall and specificity to precision. Different weight schemes have been investigated and the best results, w.r.t. recall and precision, are obtained by using term frequency with inverse document frequency and length normalization

iii. Similarity Coefficients

The similarity in vector space models is determined by using associative coefficients based on the inner product of the document vector and query vector, where word overlap indicates similarity. The inner product is usually normalized. The most popular similarity measure is the cosine coefficient, which measures the angle between the document vector and the query vector.

B) MapReduce

MapReduce is the original framework for writing applications that process large amounts of structured and unstructured data stored in the Hadoop Distributed File System (HDFS). Apache

Hadoop YARN opened Hadoop to other data processing engines that can now run alongside existing MapReduce jobs to process data in many different ways at the same time.

C) Modified K-Means:

The K-mean algorithm is a popular clustering algorithm and has its application in data mining, image segmentation, bioinformatics and many other fields. This algorithm works well with small datasets. In this paper we proposed an algorithm that works well with large datasets. Modified k-mean algorithm avoids getting into locally optimal solution in some degree, and reduces the adoption of cluster -error criterion.

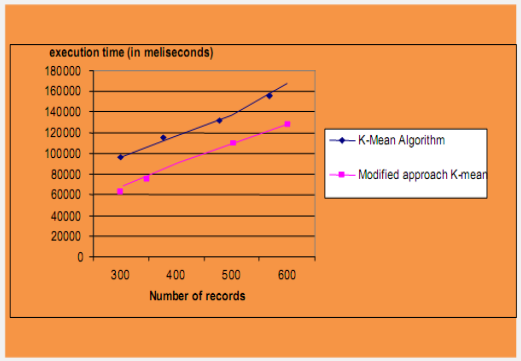
Algorithm: Modified approach (S, k), S={x1,x2,…,xn }

Input: The number of clusters k1(k1> k) and a datasetcontaining n objects(Xij+).

Output: A set of k clusters (Cij) that minimize the **Cluster** - error criterion.

Number of Records	Time taken to execute (Inmillisecond) K-Mean Algorithms	Time taken to execute (Inmillisecond)Modified K-MeanAlgorithm
300	95240	61613
400	116243	73322
500	135624	103232
600	158333	122429

Comparison between K-Mean and Modified approach algorithm with large Number of Records and its Execution Time in milliseconds is shown on the table.



Graph shows the comparison between K-mean and Modified approach K-mean on the basis of large number of records and execution time using this algorithm. Modified approach K-mean better performance in comparison to standard K-means algorithm

IV. Conclusion

Proposed model overcomes the limitations of the traditional filtering algorithm such as,

- Scalability
- Poor accuracy

Moreover, to improve the scalability and efficiency in “Big Data” environment, we have implemented it on a Map Reduce framework in Hadoop platform and Vector Space Model. In our future work, we will do further research in how to deal with the case where term appears in different categories of a domain thesaurus from context and how to distinguish the positive and negative preferences of the users from their reviews to make the predictions more accurate. The proposed system is more efficient in terms of complexity. And the system gives more accurate results or recommendations to the users.

References

[1] Maria Bielikova, Michal Kompan, and Dusan Zelenik, "Effective Hierarchical Vector-based News Representation for Personalized Recommendation", DOI 10.2298/CSIS110404070B,
[2] Shakhy.P.S, Swapna.H, "Improved Keyword Aware Service Recommendation System for Big Data Applications", International Journal of Innovative Research in Computer and Communication Engineering, Vol. 3, Issue 8, August 2015, ISSN (Print): 2320-9798,
[3] Daniar Asanov, "Algorithms and Methods in Recommender Systems",
[4] Pasquale Lops, Marco de Gemmis and Giovanni Semeraro, "Content-based Recommender Systems: State of the Art and Trends",
[5] J. Amaithi Singam and S. Srinivasan, "Optimal keyword search for recommender system in big data application", ARPN Journal of Engineering and Applied Sciences, VOL. 10, NO. 07, April 2015 ISSN 1819-6608,
[6] Ali Elkahky, Yang Song, Xiaodong He, "A Multi-View Deep Learning Approach for Cross Domain User Modeling in Recommendation Systems", International World Wide Web Conference Committee (IW3C2), May 18–22, 2015, Florence, Italy, ACM 978-1-4503-3469-3/15/05,
[7] Debajyoti Mukhopadhyay, Ruma Dutta, Anirban Kundu and Rana Dattagupta, "A Product Recommendation System using Vector Space Model and Association Rule", International Conference on Information Technology, DOI 10.1109/ICIT.2008.48
[8] Vinaya B. Savadekar, Pramod B. Gosavi, "Towards Keyword Based Recommendation System", International Journal of Science and Research (IJSR), ISSN (Online): 2319-7064, Volume 3 Issue 11, November 2014,
[9] Shunmei Meng, Wanchun Dou, Xuyun Zhang, Jinjun Chen, "A Keyword-Aware Service Recommendation Method on MapReduce for Big Data Applications", DOI 10.1109/TPDS.2013.2297117,
[10] Pallavi R. Desai, B. A. Tidke, "A Survey on Smart Service Recommendation System by Applying Map Reduce Techniques", International Journal of Science and Research (IJSR), Volume 5 Issue 1, January 2016, ISSN (Online): 2319-7064

