OPTIMAL KEYWORD SEARCH FOR RECOMMENDER SYSTEM IN BIG DATA APPLICATION

J. Amaithi Singam and S. Srinivasan
Department of Computer Science and Engineering, Regional Centre of Anna University, Madurai, Tamil Nadu, India
E-Mail: amaithi24@gmail.com

ABSTRACT
Currently, 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. Hybrid Filter algorithm generates keyword recommends from the previous user preferences. To implement effective results in big data environment, this method is implemented using the concept of Map Reduce parallel processing on Hadoop. Experimental results are shown the effective results on real-world datasets and reduce the processing time from large datasets.

Keywords: recommendation system, web search, big data, map reduce.

1. INTRODUCTION

A web search engine [9] is a software system that is designed to search for required information. The search results are generally presented in a line of results. The information may be a mix of web pages, images, document and other types of files. The input of the search engine may be a word, string or token. Data means collection of information, it may be human language or non-human language. This kind of information increases rapidly that cause’s big data problem. Big data can be categorized into volume, varity, velocity. Volume means it can process terabytes to zettabytes data. Varity may be in the form of structure, unstructured, semi structure data. Velocity means streaming of data. The big data mainly faces two problems; they are scalability problem and accuracy problem. Scalability problem [22] occurs when the amount of processing time increases while processing of large amount of data. Currently we have enormous volume of structured and unstructured data, and while searching we could not get accurate results. For example, railway ticket reservation system contains terabytes of data. So that active user finds difficult to pick the correct data.

Recommendation system depicts additional information to the active user such as learning material Recommendation system [7]. Recommendation system in real e-commerce, Music recommendation system in Apple iTunes, Amazon.com recommendation system, many more. The main role of the Recommendation system is information filtering. Recommendation system can be classified into two types, they are knowledge based recommendation and collaborative filtering [9][10]. The widely used technique is collaborative filtering. The collaborative filter only focuses on the user preference of active user. Knowledge based recommendation system means asking a user about their own preference of the item and find out what active user required.


Demographic filter provides the additional information to the required item. The widely used another filter is content based filter [21]. The main work of the content based filter is item oriented similarity, it provides widely desired item in the past. Collaborative filter (CF): is most effective filter in recommendation system. The collaborative filter depends on the user preference, user interest or user taste recommendation. The collaborative filter automatically appends the up-to-date information. Compared with content based filter it only concentrates on the item but does not focus on user preference. CF [14][17] can be classified into two types, they are memory based and modern based. Memory based is based on the database, also used item to item or user to user based filtering techniques it fully depend on the correlation based techniques to predict the future item. A model based technique does not use whole data set to predict the item. The most widely used techniques in model based are clustering based techniques and single value decomposition techniques.

Our proposed work is optimal keyword search for recommendation system in big data application. Here we use Hybrid filter. Hybrid filter is a combination of both content and CF techniques. Here we consider Hotel reservation system that provides best hotel in preferred area based on the user request. Finally we compare the efficiency and scalability [23] for different datasets.

2. RELATED WORK

The service recommendation [1] system provides additional information to the user. Currently the amount of data has increased rapidly that yields big data problem. In traditional service method faces two problems when processing large amount of data. They are efficiency of
data and scalability problem. Moreover most of the traditional recommended systems focus on the ranking and ratings of service with different services and different user. Existing system does not consider the user preference [2], only focuses on the item of the service. Shunmei Meng et.al. (2013) proposed the service recommendation system [5] used the collaborative filter algorithm to generate the approximate recommendation to the active user. That improves the scalability and efficiency of the data when processing large amount of data. Works are implemented on Hadoop using MapReduce [12] framework with real time dataset. 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 [6] 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 [19] 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 [3] 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 viewer’s choices on the product advertisement are important part of the market. This part generates the recommendation [11] [13] 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 recommendation system in an e-learning context [4] is efforts to intelligently recommend to the active user (learner) based on the action of the previous user (learner). Also this recommendation system is based on the online activity such as reading posted message on internet. These recommendation systems have been tried in an e-commerce [13] to purchase the quality goods. Dhoha Almazzo et al (2013) addresses the use of web mining to build the online activities based on user (learner) access History to improve course material navigation as well as improve the processing time. Also e-learning compared to the traditional system of face-to-face style teaching and learning. It provides more benefits than the face-to-face learning. Here the content based filter is used. The content based filter to use the algorithm of vector space model for similarity computation to predict the similar item based on the user rated item that improve the accuracy of data.

The information about the product is increasing with exponential rate in e-commerce industry [15]. This environment has complexity to find the optimal information in majority of data collection. The customer can get benefit by receiving optimal information about the products which they are customer likely to buy. And discussed about the different type of recommendation system filters [16] method and mainly focused on the algorithm of collaborative filter and comparing each and every approaches of collaborative filter. They are 1.Active filter, 2.Passive filter, 3.Item based filter. The Active filter uses the peer to peer approaches. Passive filter focuses on the implicit collection of information. The item based filter only addresses the item of the user preference. Jong Seo Lee (2012) describes the different kind of recommendation system. They are 1.Musical recommendation system in apple iTunes, 2.Recommendation system in amazon, 3.Recommendation system in like-i-like, 4.Music Surfer beyond Existing Music Recommendation system. And also describe the following approaches, they are 1. Clustering model, 2.Bayesian network model, 3.regression approaches.

Bigtable [18] stored the structure data in distributed storage system manner to scale the very large size. Currently, it can store beyond the terabytes of the data. Google store, Web indexing, Google earth, Google Finance also used a BigTable for storing structured data. The Bigtable has successfully provided a flexible, high performance solution for the entire Google product. F. Chang et al. 2008 described the data model such as sparse, distributed and persistent multidimensional data. All data models are stored in rows and column manner. And used analogous of three level hierarchies and evaluates the performance of data storing and recovery compared to the traditional methods.

3. ARCHITECTURE

We gathered the existing user’s reviews and stored it into the dataset in Figure-1. The Data set was downloaded from UCI repository. It is named as Hotel Reservation System. The attributes of this dataset are Hotel name, Date of the Review and Users review. The preprocessing step is used to remove the noise from dataset. After removing the noise the data is stored into database.
These processes are under the offline. Then the Active user gives their preferred item and rated of the item. That preferred item was compared to the previous user preference. If the Active user preferred item is similar to the previous user item then the required similar item is rated and ranked. Finally, we generate the Recommendation to the Active user. Online environment similarity checking, Ranking and Rating, Recommendation Generation are performed using map reduce concept.

Preprocessing
The previous user’s reviews are stored into the Dataset. The Dataset was downloaded from the UCI repository. The size of the Dataset is 257MB. The Dataset was stored in the text files. Main attributes of the Dataset is Hotel Name, Year, Previous user Reviews.

The main function of preprocessing is to remove the noise from the previous user reviews and required keywords are extracted from the data set using porter stemmer algorithm.

Porter stemmer algorithm
The porter stemmer algorithm is used to remove the suffix word from the previous users review. And the common morphological also elite that term called stem. The document represents the term or vector form. Here the previous user review keywords are classified using porter stemmer algorithm.

Previous user review (PUR) keyword1= neatness
Previous user review (PUR) keyword2= neat

**Step-1:** compare {(PURK1), (PURK2)}
**Step-2:** REMOVE suffix → ness
**Step-3:** Stem = neat

Similarity checking
The previous user reviews are extracted from Dataset and stored the database. The attributes are Hotel name, Date, keywords, count. The active user gives their required Keywords with rating. The previous user’s keywords (PUK) are compared to the Active user’s keywords (AUK). If the keywords are exactly same then the keywords are stored in the matrix using Pearson correlation coefficient algorithm this method is called exact similarity. Otherwise the method is said to be approximate similarity.

**Approximate similarity method**
Set theorem is used to find out the similarity between the active user keywords and the previous user keyword.

\[ \text{Sim} (\text{AUK}, \text{PUK}) = \frac{\text{AUK} \cap \text{PUK}}{\text{AUK} \cup \text{PUK}} \quad (1) \]

**Exactly similarity (Collaborative filter)**
The required keywords of the active user and previous user keywords are transformed into the form of n dimensional weight vector. And similarity and weight was calculated using Pearson correlation coefficient algorithm.

**Pearson correlation coefficient algorithm**
Correlation between the set of keywords is a measure of how well they are related to the similar. Pearson Correlation Coefficient Algorithm is used to measures the linear relationship between two sets of Keywords and also measures the user preference.

\[ \text{Sim} (\text{a,p}) = \frac{\sum_{k \in K} (r_{a,k} - r_{a}) (r_{p,k} - r_{p})}{\sqrt{\sum_{k \in K} (r_{a,k} - r_{a})^2} \sqrt{\sum_{k \in K} (r_{p,k} - r_{p})^2}} \quad (2) \]

where K is the set of Keywords rated by both users, \( r_{a,k} \) and \( r_{p,k} \). \( r_{a} \) and \( r_{p} \) denoted the rating of both active users end previous users. \( r_{a} \) and \( r_{p} \) represents the mean value of active user and previous user. This is under the Collaborative filter.

\[ W (k1,k2) = \frac{\sum_{u \in U} (r_{u,k1} - r_{1}) (r_{u,k2} - r_{2})}{\sqrt{\sum_{u \in U} (r_{u,k1} - r_{1})^2} \sqrt{\sum_{u \in U} (r_{u,k2} - r_{2})^2}} \quad (3) \]

where \( k1, k2 \) keywords are selected by user’s, here U denotes the set of Users, \( r_{i,1} \) and \( r_{i,2} \) represents the mean value of the keywords k1 and k2.

**Rating and ranking (content based filter)**
The rating value can be calculated by the similar users. Then the service recommendation list will be generated from that user can collect the highest rating service using nearest neighbor’s algorithm.

**Top k with nearest neighbor’s algorithm**

\[ G_{a,k} = \frac{r_{a} + \sum_{k \in K} (r_{a,k} - r_{a}) \times W(k1,k2)}{\sqrt{\sum_{k \in K} (r_{a,k} - r_{a})^2} \times W(k1, k2)} \quad (4) \]
\[ \sum_{a,k} W(k_1,k_2) \]

Where \( G_{a,k} \) is the prediction for the active user \( a \) for keyword \( k \), \( W(k_1,k_2) \) is the similarity between users \( a \) and \( p \), and \( K \) is the neighborhood or set of most similar users.

Map Reduce

Map Reduce [1] [3] [23] is used to execute the data in parallel manner. Here the similarity checking, weight generation, rating and ranking are executed in parallel [20].

4. Experimental Results

This section describes the performance analysis to validate the proposed algorithm.

Accuracy

Experimental results demonstrate the efficiency of the proposed Pearson correlation algorithm. Figure-2 shows the comparison of cosines theorem and the Pearson correlation coefficient algorithm.

Figure-2. Algorithm comparison.

Figure-2 depicts the comparison of Cosine and Pearson correlation. For keyword 3 our proposed algorithm shows a probability of 0.6 whereas cosine based algorithm shows the probability of 0.4. Our proposed algorithm increases the probability 0.2 when we compared to cosine based algorithm.

Scalability

When the data set is increased, we have scalability problem.

Figure-3. Scalability checking.

Figure-3 depicts the comparison of Collaborative filter and hybrid filter. For Count 3 our proposed filter shows a processing time of 0.2141ms whereas Collaborative filter shows processing time of 0.3172ms. Our proposed filter reduces the processing time upto 0.1031ms when compared to Collaborative filter.

5. Conclusions and Future Enhancement

We have proposed a Service Recommendation System on Map Reduce concept. In this work keywords are used to indicate users’ (active and existing) preferences, and a hybrid Filtering algorithm is adopted to generate appropriate recommendations. The active user gives the preferences by selecting the keywords with rating from the keyword candidate list. The preferences of the previous users can be extracted from their reviews for services corresponding to the keyword-candidate list. Our method aims to improve the scalability and efficiency of Big Data environment, Implemented on map reduce concept.

References


