

User Based Personalized Search With Big Data

S. EbenazerRoselin^[1], K. Anandhi^[2], K. Durairaj^[3], R. Hariharan^[4],
Member IEEE^{[1][2]} PG scholar, ^{[3][4]} Assistant Professor
^{[1][2][3][4]} Department of Information Technology,
Vel Tech University, Chennai, India
^[1] ebenazerroselin89@gmail.com, ^[2] aniweb10@gmail.com
^[3] durait2011@gmail.com, ^[4] hharanbtech@gmail.com

Abstract

The data usage now is getting enhanced rapidly leading to excessive area desired for it to conserve. So in Big data to address this with its inproficiency problem and ability to its accommodation of growth with inherited support proposing technique which already fails to meet personalized necessity and various inclination, the user based personalized search with big data is used by applying keyword aware support processing techniques. Here, we first use keywords taken from rankings of previous end user about their desires. The UBCF algorithm is adopted to generation of appropriate recommendations. We also bring it on a distributed computing platform, Hadoop which uses map reducing as its computing frame work

Keywords: Collaborative Filtering, Big Data, Map Reduce, Hadoop

Introduction

The works done on the data in last few years has been enhanced so much which give arise to the word so as called “Big Data” which creates an interest to work on with the data which have their size beyond the current technology, various methods to store and its processing and working time. The researchers are made in such a way that makes the effectiveness as its major criteria. Here the support processing technique is used for giving various suggestions to the customers. Already done works cannot overcome the in proficiency problems and ability to make up with the current technology problems while processing more amount of data is used. They also provide the same ranking procedure which actually fails to meet the customer preferences.

The current referencing methods are usually classified into three major categories such as content-based, collaborative and hybrid recommendation approaches. In this the content based and collaborative based approaches they refers to some works that

are done in past. The hybrid approaches combines both the works in different ways. The collaborative filtering sub divides the process as item based collaborative filtering and user based collaborative filtering.

Literature Survey

The existing service recommender systems present the same ratings and rankings of services to different users. It proposes a Keyword-Aware Service Recommendation method, named KASR, to address the challenges such as scalability and inefficiency problems when processing or analysing a large-scale data.[1]. The suggestions in this paper relates to the techniques applied for Personal Program Guide named PPG. The PPG manages a user model for each individual user's preferences for TV program categories. It becomes easy as the PPG runs on the set-top box and strongly integrated with the TV playing and the video recording services offered by that [6]. The channel scheduling is done by provided content which is more interesting for the people for their gain. It is based on the content uploaded by the people, properly classified. It is done with the recommended systems for Audio visual contents with special considerations for the group of viewers. [2]. Many projects in Google store data in Big table, such as Google earth, web indexing and Google finance cross the data across thousands of commodity servers. So these plays a vital role in demands on Big table both in data size and latency requirements. It provides the simple data model provided by Big table. [3]. For e-commerce websites the usage of Recommendations algorithms is one of the simplest technique to generate a list of suggestions. Many projects uses to rate only their interests. But we can use other ratings such as including item viewed, demographic data, subject interest and favorite artists. This paper recommends the online store for customers. They have interests based on customer interests, showing programming titles to a software engineer. E-commerce recommendation algorithms often operative in challenging positions such as large and huge amounts of data , high quality suggestion, differentiation between old and new customers and volatile data. This algorithm produces suggestion in real time, scales to massive data sets and generates high quality suggestion. [5]. finally the user travel suggestions and to show splendid results. It conducts special travel suggestions by considering user profiles and attributes. By verifying the association of people attribute such as time, popular landmarks etc. It also works to show the nearby locations in mobile using android applications. [4] The researches on the process management gets increased but the toughest job is to conduct process analysis due to huge amount of data. So by this project we collect the event logs called big data and analyze the collected logs with the processes as structured and unstructured data. Usually only structured processes are identified but we also use unstructured data for enhanced clarity results. The system automatically discovers a process model and conducts various performance analysis on the manufacturing processes.[7] The knowledge of dynamic traffic and its usage of data services in cellular networks is important for network resources and improving users experience. The studies related on its behavior, device type and application is made. Here we use service providers, to reveal the traffic characters in cellular data networks. Results of our study present

mobile Internet participants with a better understanding of the traffic and usage characteristics of service providers, which play a critical role in the mobile Internet era. [8]. Here it suggests various methods for catering to the problems in hand through Map Reduce framework over Hadoop Distributed File System. Map Reduce is a Minimization technique which makes use of file indexing with mapping, sorting, shuffling and finally reducing. Map Reduce techniques have been studied in this paper which is implemented for Big Data analysis using Hadoop Distributed File System. [9]. This paper reports the experimental work on big data problem and its optimal solution using Hadoop cluster, Hadoop Distributed File System (HDFS) for storage and using parallel processing to process large data sets using Map Reduce programming framework. We have done prototype implementation of Hadoop cluster, HDFS storage and Map Reduce framework for processing large data sets by considering prototype of big data application scenarios. The results obtained from various experiments indicate favorable results of above approach to address big data problem. [10] To reduce the computation and to focus the mining for the latter situations, we propose a tree-based algorithm that (i) allows users to express the patterns to be mined according to their intention via the use of constraints and (ii) uses MapReduce to mine uncertain Big data for only those frequent patterns that satisfy user-specified constraints. Experimental results show the effectiveness of our algorithm in mining interesting patterns from uncertain Big data.[11] Spatial data is one of the major part in big data, governing people interests. Spatial data mining confronts much difficulty when attracting the value hidden in spatial big data. This paper helps in techniques to discover knowledge from spatial big data that may help data to become intelligent. [12].

Related Work

1. Most of the existing researches such as hotel reservations or hospitals, the rating services are only followed and the lists given to them are same. They have no preference choices for the users where it fails to meet the customer needs and preferences.
2. Also the another major drawback is that many services are quoted into a single checkbox rating where the working of many interior work is almost not known to the customers.
3. The existing researches tries to solve the scalability problem by dividing the dataset and to evaluate them separately and bringing it together. But here also it becomes a quite difficult process once if the large amount of data are used.

User Based Personalized Search

In this paper User Based Collaborative Filtering Algorithm is used. Here the keywords are filtered from the last used customers to show their preferences. Also we use it on a distributed computed platform, Hadoop which uses map reduce as its computing framework.

Here, the keywords are used to indicate both of customers ideas and the quality of their services. The user based collaborative filtering algorithm is adopted to generate their preferences. In the fig:1 the goal is for making a calculation of a personalized rating at each of their services done for the user and then by giving a personalized service recommendation list and give the most useful information based upon their need. Then also to enhance the scalable property and efficient of our service method in “Big Data” environment, we implement it on a map reduce framework on hadoop by splitting the proposed algorithm in multiple map reducing phases.

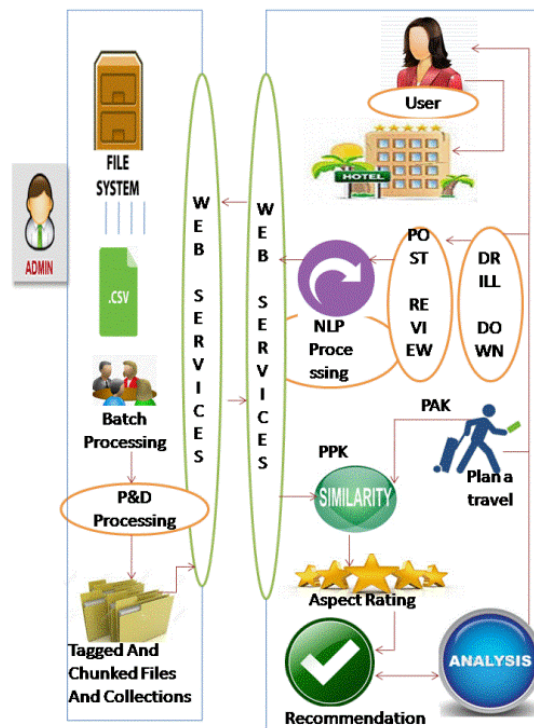


Figure 1: User Based Personalize Search

Big Data Collection and Environment

The very large amount of data is retrieved from the open source data sets which are available from the travel recommended applications. Big data schemas are analyzed and the working set of the program is defined. The CSV(Comma separated values) files were read and manipulated using Java API in which itself developed by us in which the developer friendly, light weighted and easily editable.

Mining in Big Data

The CSV Files in distributed Systems are invoked through Web Service Running in the Server Machine of the Host Process through a Web Service Client Process in the Recommendation System. The data that Retrieved to the Recommendation Systems are provided with a clean GUI and can be Queried on Demand. Each and Every

process on the Recommendation Application invokes Web Service which uses light weighted traversal of data using XML. The Users can Review each hotel and can post comments also. The Reviews gets updated to the CSV Files as it get retrieved.

Service Recommender Application

The Traditional View of Service Recommender Systems that shows Top-K Results are displayed with Paginations with which a user can navigate Back and Forth of the Result sets. All Services Ratings and Reviews of Each Hotels are Listed. A User can Plan or Schedule a Travel highlighting his requirements in a detailed way that shows the Preference Keywords Set of the Active User. A Domain Thesaurus is Built depending on the Keyword Candidate List and Candidate Services List. The Domain Thesaurus can be Updated Regularly to get accurate Results of the Recommendation System.

Map Reduce and Hadoop

Capture user preferences by a keyword-aware approach:

In this step, the preferences of active users and previous users are formalized into their corresponding preference keyword sets respectively. In this paper, an active user refers to a current user needs recommendation.

a) Preferences of an active user.

An active user can give their ideas about user need services by selecting keywords from a keyword-user list, which reflect the quality criteria of the services they concerned about. Besides, the active user should also select the very needed of the keywords. The importance degree of the keywords represents the general “3” represents important and “5” represents very important.

b) Preferences of previous users.

The preferences of last users for a candidate service are extracted from their reviews for the service according to the keyword-candidate list and domain thesaurus. And a review of the previous user will be formalized into the preference key-word set of User.

The keyword extraction process is described as follows:

a) Preprocess:

At First, the HTML tags and stop words in the reviews snippet collection should be removed to avoid affecting the quality of the keyword extraction in the next stage. And the Porter Stemmer algorithm is used to remove the commoner morphological and inflexional endings from words in English.

b) Keyword extraction:

In this phase, each review will be transformed into a corresponding keyword set according to the keyword-candidate list and domain thesaurus. If the review contains

a word in the domain thesaurus, then the corresponding keyword should be extracted into the preference keyword set of the user

Similarity Computation

The Third step is to identify the reviews of previous users who have similar tastes to an active user by finding neighborhoods of the active user based on the similarity of their preferences. Before similarity computation, the reviews unrelated to the active user's preferences will be filtered out by the intersection concept in set theory. If the intersection of the preference keyword sets of the active user and a previous user is an empty set, then the preference keyword set of the previous user will be filtered out.

Methodology

Basic Algorithm for User Based Personalized Search With Big Data

Input: The preference keyword set of the active user KSW

The candidate services $WS = \{ws1, ws2, \dots, wsN\}$

The threshold δ in the filtering phase

The number K

Output: The services with the Top-K highest ratings $\{tws1, tws2, \dots, twsK\}$

1: for each service $ws_i \in WS$

2: $\hat{R} = \phi, sum = 0, r = 0$

3: for each review R_j of service ws_i

4: process the review into a preference keyword set PSW_j

5: if $PSW_j \cap KSW \neq \phi$

then

6: insert PSW_j into \hat{R}

7: end if

8: end for

9: for each keyword set $PSW_j \in \hat{R}$

10: $sim(KSW, PSW_j) = SIM(KSW, PSW_j)$

// $SIM(KSW, PSW_j)$ can be $SIM-ASC(KSW, PSW_j)$ or $SIM-ESC(KSW, PSW_j)$

11: if $sim(KSW, PSW_j) < \delta$

then

12: remove PSW_j from \hat{R}

13: else $sum = sum + 1, r = r + r_j$

14: end if

15: end for

16: $\bar{r} = r / sum$

17: get pr_i by formula (7)

18: end for

19: sort the services according to the personalized ratings pr_i

20: return the services with the Top-K highest ratings

$\{tws1, tws2, \dots, twsK\}$

Exact Similarity Computation

Input: The preference keyword set of the active user KSW

The preference keyword set of a previous user PSW_j

Output: The similarity of KSW and PSW_j , $simESC(KSW, PSW_j)$

- 1: for each keyword k_i in the keyword-candidate list
- 2: if $k_i \in KSW$ then
- 3: get $\vec{W}_{AP,i}$ by formula (2)
- 4: else $\vec{W}_{AP,i} = 0$
- 5: end if
- 6: if $k_i \in PSW_j$ then
- 7: get $\vec{W}_{PPj,i}$ by formula (5)
- 8: else $\vec{W}_{PPj,i} = 0$
- 9: end if
- 10: end for
- 11: get $simESC(KSW, UPK_j)$ by formula (6)
- 12: return the similarity of KSW and PSW_j , $simESC(KSW, UPK_j)$

Input: The preference keyword set of the active user ASW

The preference keyword set of a previous user PSW_j

Output: The similarity of ASW and PSW_j , $simASC(ASW, PSW_j)$

$$\begin{aligned}
 1. \quad sim_{ASC}(KSW, PSW_j) &= \frac{|KSW \cap PSW_j|}{|KSW \cup PSW_j|} \\
 2. \quad sim(KSW, PSW) &= \cos(\vec{W}_{AP}, \vec{W}_{PP}) = \frac{\vec{W}_{AP} \cdot \vec{W}_{PP}}{\|\vec{W}_{AP}\|_2 \cdot \|\vec{W}_{PP}\|_2} \\
 &= \frac{\sum_{i=1}^n \vec{W}_{AP,i} \cdot \vec{W}_{PP,i}}{\sqrt{\sum_{i=1}^n \vec{W}_{AP,i}^2} \sqrt{\sum_{i=1}^n \vec{W}_{PP,i}^2}}
 \end{aligned}$$

Performance Evaluation

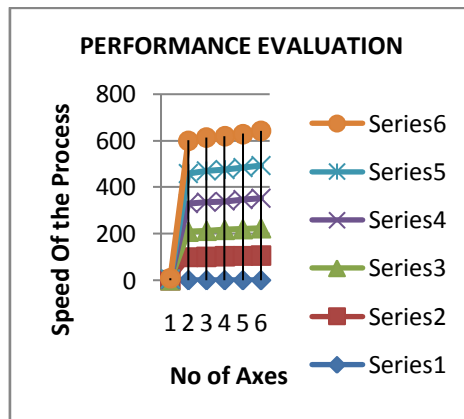


Figure 2: Performance Evaluation

To verify the scalable performance fig:2 of the project, an experiment is done respectively in a cluster of nodes ranging from 1 to 8. There are 4 synthetic datasets used in the experiments (128M, 256M, 512M and 1G datasize). Figure shows the speedup of User Based Personalized Search With Big Data . From Figure, it shows speedup of User Based Personalized Search With Big Data increases relative linearly with the growth of the number of nodes. Meanwhile, larger dataset obtained a better speedup. When the datasize is 1G and the number of nodes is 8, the speedup value reaches 6.412, which is 80.15% ($6.412/8=80.15\%$) of the ideal speedup. The experimental result shows that User Based Personalized Search With Big Data on Map-Reduce in Hadoop platform has good scalability over “Big Data” and performs better with larger dataset.

Enhancement

The Natural Language Processing is implemented to analyze the reviews of the previous user. The NLP Process Comprises Tokenizing a Sentence or a word ,POS (Parts of Speech) Tagging ,Extraction of Nouns and Verbs, Synonym Retrieval of Extracted Keywords using WordNet Dictionary .The BigData manipulations from CSV through Our Own JAVA API enforces developer friendly access.

Conclusions and Future Work

In this paper, we have proposed the user based personalized search with big data is used by applying keyword aware support processing techniques.. Here, key-words are used to indicate users' preferences with the positive and negative preferences of the users from their reviews to make the predictions more accurate and a user-based Collaborative Filtering algorithm is given for recommendations. Moreover, a keyword-candidate list and domain thesaurus are provided to help obtain users' preferences. The active user gives his/her preferences by selecting the keywords from the keyword-candidate list, and the preferences of the previous users can be extracted from their reviews for services according to the keyword-candidate list and domain thesaurus. Our method aims at presenting a personalized service recommendation list and recommending the most accurate services to the users. Moreover, to improve the scalability and efficiency of this system in “Big Data” environment, we have implemented it on a MapReduce framework in Hadoop platform. Finally, the experimental results demonstrate that based personalized search with big data significantly improves the accuracy and scalability of service recommender systems over existing approaches.

References:

- [1] Shunmei Meng, Wanchun Dou, Xuyun Zhang, Jinjun Chen KASR: A Keyword-Aware Service Recommendation Method on MapReduce for Big

- Data Applications. In Proceedings of IEEE TRANSACTIONS ON PARALLEL AND DISTRIBUTED SYSTEMS, TPDS-2013-12-1141
- [2] Rafael Sotelo, IEEE Member, Jose Joskowicz Alberto Gil Solla - An affordable and inclusive system to provide interesting contents to DTV using Recommender Systems, 2012 IEEE International Symposium on Broadband Multimedia Systems and Broadcasting
 - [3] Fay Chang, Jeffrey Dean, Sanjay Ghemawat, Wilson C. Hsieh, Deborah A. Wallach
 - [4] Mike Burrows, Tushar Chandra, Andrew Fikes, Robert E. Gruber-Bigtable:
 - [5] A Distributed Storage System for Structured Data, ffay, jeff, sanjay, wilsonh, kerr, m3b, tushar,_kes,gruberg@google.com
 - [6] Google, Inc.
 - [7] S.Saranya, Student of Information Technology, A. Ramachandran, Assistant Professor of Information Technology, Individualized Travel Recommendation by Mining People Ascribes and Travel Logs Types from Community Imparted Pictures in S.Saranya et al, / (IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 5 (2) , 2014, 1685-1687
 - [8] Greg Linden, Brent Smith, and Jeremy York • Amazon.com, Amazon.com Recommendations Item-to-Item Collaborative Filtering, Industry Report
 - [9] Liliana Ardissono, Cristina Gena, Pietro Torasso Dipartimento di Informatica, Università di Torino, Corso Svizzera 185, 10149Torino, Italy, Fabio Bellifemine, Angelo Difino, Barbara Negro Telecom Italia Lab, Multimedia Division, Via G. Reiss Romoli 274, 10148 Torino, Italy- User Modeling and Recommendation Techniques for Personalized Electronic Program Guides
 - [10] Liliana Ardissono, Cristina Gena, Pietro Torasso Dipartimento di Informatica, Università di Torino, Corso Svizzera 185, 10149Torino, Italy, Fabio Bellifemine, Angelo Difino, Barbara Negro Telecom Italia Lab, Multimedia Division, Via G. Reiss Romoli 274, 10148 Torino, Italy- User Modeling and Recommendation Techniques for Personalized Electronic Program Guides
 - [11] Hanna Yang ; Sch. of Bus. Adm., Ulsan Nat. Inst. of Sci. & Technol., Ulsan, South Korea ; Minjeong Park ; Minsu Cho ; Minseok Song , A system architecture for manufacturing process analysis based on big data and process mining techniques , IEEE in Washington, DC On10.1109/BigData.2014.7004336
 - [12] Liu Jun ; Beijing Key Lab. of Network Syst. Archit. & Convergence, Beijing Univ. of Posts & Telecommun., Beijing, China ; Li Tingting ; Cheng Gang ; Yu Hua more authors , Mining and modelling the dynamic patterns of service providers in cellular data network based on big data analysis- IEEE On Dec. 2013

- [13] Manikandan, S.G. ; Dept. of Inf. Technol., Dhanalakshmi Coll. of Eng., Chennai, India ; Ravi, S. Big Data Analysis Using Apache Hadoop in IEEE Conference
- [14] Patel, A.B. ; Birla, M. ; Nair, U. Addressing big data problem using Hadoop and Map Reduce in IEEE Conference
- [15] Leung, C.K.-S. ; Dept. of Comput. Sci., Univ. of Manitoba, Winnipeg, MB, Canada ; Fan Jiang , A Data Science Solution for Mining Interesting Patterns from Uncertain Big Data in IEEE Conference
- [16] Shuliang Wang ; Sch. of Software, Beijing Inst. of Technol., Beijing, China ; Hanning Yuan Spatial Data Mining in the Context of Big Data in IEEE Conference