International Journal of Science and Research (IJSR)
ISSN (Online): 2319-7064
Index Copernicus Value (2013): 6.14 | Impact Factor (2013): 4.438

Provision of Content Based Service Recommendations using Hadoop and MapReduce

M. Vigneesh[1], K. Nimala[2]

[1]M.Tech Cloud Computing, Department of Information Technology
SRM University, SRM Nagar, Kattankulathur, Tamil Nadu, India, 603203

[2]Assistant Professor Sr.G, Department of Information Technology
SRM University, SRM Nagar, Kattankulathur, Tamil Nadu, India, 603203

Abstract: ***Service recommender systems are widely known as the valuable tools for providing appropriate recommendations to users. Most of the existing service recommender systems give the same ratings and rankings of services to different users without considering diverse users' preferences, therefore it fail to meet users' personalized requirements. In recent years the amount of customers, services and online information has grown rapidly, yielding the big data analysis problem for service recommender systems. Traditional service recommender systems face performance issues like scalability and inefficiency problems when processing or analyzing such large-scale data. Here, we propose a method to address the above issues which aims at pre sending a personalized service recommendation list and recommending the most appropriate services to the users with different perspective. Specifically, keywords are the most helpful things used to indicate users' diverse preferences, and a user-based Collaborative Filtering algorithm is used to generate appropriate recommendations. To improve scalability and efficiency in a big data environment, the proposed system is implemented on Hadoop platform, which is a widely-adopted distributed computing platform using the MapReduce parallel processing paradigm. Hence, the proposed system significantly improves the accuracy and scalability of service recommender systems and provides the recommendations to the users efficiently.***

Keywords: Hotels, Recommender Systems, Hadoop, Map Reduce

1. Introduction

In recent years the key concept of Big Data comes into picture since the amount of data in our world has been increasing in larger size. Big Data is a popular term which refers to larger datasets where size is beyond the ability of current technology and also it is a biggest challenge in IT industry. There are some alternative services, which are efficiently recommending services that users preferred has been still a research part of service recommender systems. Because Service recommender systems are prevalent as the valuable tools to help users to deal with services overload and provide some appropriate recommendations to them. Service Recommender systems are defined as a system which produces individualized recommendations as the output and act as the personal guide to recommend services.

Currently, the available recommendation methods can be characterized into three main categories: content-based, collaborative, and hybrid recommendation approaches. Content-based approaches recommend services similar to those the user preferred in the past. Collaborative filtering (CF) approaches are found as the useful method to recommend services to the users with similar tastes preferred in the past. Hybrid approaches combine content-based and CF methods in several different ways. Collaborative Filtering (CF) is a classic personalized algorithm which is widely used in many commercial recommender systems [13]. In CF based systems, users receive recommendations based on people who have similar tastes and preferences, which can be further classified into item-based CF and user-based CF. In item-based systems, the predicted rating depends on the ratings of other similar items by the same user. While in user-based systems, the predicted rating depends upon the ratings of the same item by similar users in the past.

Here, the proposed methodology will take advantage of a user-based CF algorithm to deal with user preferences and predicted ratings. Cloud Computing is the best paradigm for implementing service-oriented computation. There are some cloud computing tools like Hadoop, Mahout and MapReduce. Hadoop is the most popular and widely used open source platform inspired by MapReduce programming framework and can be used to improve the efficiency and scalability of Big Data environment.

2. Literature Survey

2.1 An affordable and inclusive system to provide interesting contents to DTV using Recommender Systems

An affordable system that feeds inclusive DTV with content interesting to the public, giving citizens new opportunities to express themselves in society is proposed in this project. It aims to two objectives: (i) to facilitate the developing by providing interesting content, and (ii) to stimulate the participation of the people in the generation of content. The core of the contribution is in the recommendation agent and the preparation of the schedule emission. The audio visual contents can be uploaded by the citizens either from their home or from an internet café. In order to create metadata for this content, certain information is requested from the user, to classify the clip in TV ontology, plus some relevant information such as author or location. This approach is based on the AVATAR recommender system which has been implemented for individuals. AVATAR stores user profiles, built based on user interests in different classes of

the hierarchies of the TV ontology, inferred from his qualifications in certain clips offered and his viewing history. Based on user profile and the metadata of a proposed content, AVATAR computes the *Predicted Degree of Interest (*PDOI) of the user for the content. PDOI is an estimation of the interest the user will have for the proposed content. Using AVATAR a list of PDOIs of each virtual user for the content available is generated. Clips with higher rankings will be those of higher interest for each virtual user. The algorithm that computes the programming of each channel is based on this ranking, while weighting the distribution list of viewers throughout the day and week. The viewer can make a login from a Set Top Box, so the system knows in real time those who are watching television. The AVATAR framework originally included a multidimensional ontology about the domain of television where classes represented the concepts of television. It is designed considering some necessities from developing countries. Because of the usage of service recommender systems the viewers would be able to get apt suggestions but the application suffers from scalability issues. If there is a sudden increase in the number of subscribers then the application suffers to handle the load. The method proposed in this project however addresses this scalability issue with the use of a Hadoop and Map Reduce environment.

2.2 Amazon.com recommendation algorithms – Item to item collaborative filtering

Recommendation algorithms are best known for their use on e-commerce websites where they use input about a customer's interests to generate a list of recommended items. Many applications use only the items customer purchases and explicitly rate to represent their interests, but they can also use other attributes, including items viewed, demographic data, subject interests and favorite artists. Amazon.com uses recommended algorithms to personalize the online store for each customer. The store radically changes based on customer interests, showing programming titles to a software engineer and baby toys to a new mother. The click-through and conversion rates are the two important measures of web based and email advertising effectiveness – vastly exceed those of untargeted content such as banner advertisements and top-seller lists. There are 3 common approaches to solving the recommender problem. Traditional collaborative filtering, cluster models and search based methods are the three methods generally followed. Most recommender algorithms start by finding a set of customers whose purchased and rated items overlap the users' purchased and rated items. The algorithm aggregates items from these similar customers, eliminates items the user has already purchased or rated and recommends the remaining items to the user. Two popular versions of these algorithms are collaborative filtering and cluster models. A traditional collaborative filtering algorithm represents a customer as an N-dimensional vector of items, where N is the number of distinct catalog items. The components of the vector are positive for purchased or positively rated items and negative for negatively rated items. To compensate the best-selling items, the algorithm typically multiplies the vector components by the inverse frequency, making less well-known items much more relevant. The algorithm generates recommendations based on a few customers who are most similar to the user. It can measure the similarity between 2 customers. Amazon.com uses recommendations as a targeted marketing tool in many email campaigns and on most websites' pages, including the high traffic amazon.com webpage. Here they use item to item collaborative filtering which mainly focuses on the items and its aim is to provide overwhelming number of suggestions. This project on the other hand uses user based collaborative filtering which mainly focuses on the number of users and provides minimum number of results which are tailor made for the users.

2.3 Individualized Travel Recommendations by Mining People Ascribes and Travel Logs Types from Community Imparted Pictures

A probabilistic personalized travel recommendation model considering users' attributes as well as their group types and the knowledge mined from travel logs is proposed in this paper. The association of a person's attributes and more contexts are investigated (e.g., time, popular landmarks) and the benefits for profiling human activities are shown. Here a social networking kind of website is developed focusing mainly on collecting users' photos. The data are mined whenever needed in order to provide suggestions. The technique used here is data mining and when compared to using a big data and map reduce framework this is not reliable.

3. Problem Definition

The traditional Service Recommendation Systems provide recommendations based on the preferences and not on the user requirements. Hence it results in the recommendations not being personalized on user recommendations. Moreover, the recommendations provided to the user are based on numerical ratings. The other challenge of this system is handling efficiency at large volumes of data. There is no provision of scalability methods for growing data.

4. Proposed System

The main objective of the proposed system is to develop a recommender system which addresses scalability and inefficiency problems in big data with the traditional service recommender systems which fail to meet the users' personalized requirements and diverse preferences. The proposed system architecture is shown Fig 1.

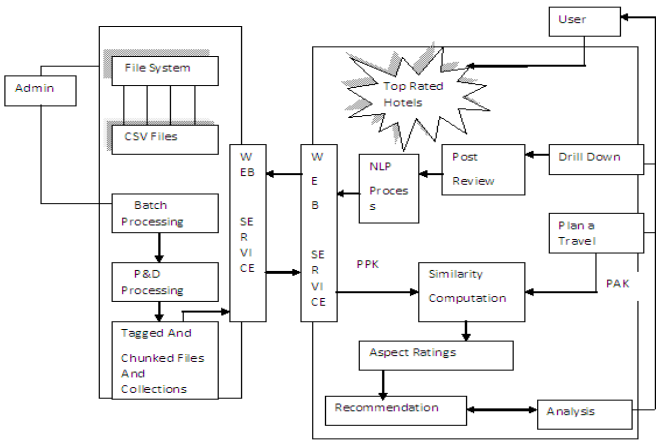

Figure 1: System Architecture

4.1 Big data Environment

The Big Data Environment has large-scale data sets which have been used while analyzing and processing real time recommendation applications such as travel recommendation applications. It has a huge collection of data which are retrieved from open source datasets that are publicly available from major travel recommendation applications. The Big Data Schemas were analyzed and a Working Rule of the schema is determined. The CSV (Comma separated values) files containing the data are read and manipulated using Java API which has properties such as developer friendly, light weight and easily modifiable.

4.2 Data Communication with Application

The CSV Files in distributed systems are invoked through web services running in the server machine. The data 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 given by the user gets frequently updated to the CSV Files as it gets retrieved.

4.3 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 the 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 is updated regularly to get accurate results of the recommendation system.

4.4 Hadoop and MapReduce

4.4.1 Capture user preferences by a keyword-aware approach

(a) Preferences of the current user

A user who is currently in need of service, can give his/her preferences about candidate services by selecting keywords from a keyword-candidate list, which reflect the quality criteria of the services he/she is concerned about. Besides, the current user should also select the importance degree of the keywords**.** The importance degree of the keywords is shown as 1 to represent the general, 3 to represent important and 5 to represent very important.

(b) Preferences of the previous user

The preferences of a previous user for a candidate service are extracted from his/her reviews for the service according to the keyword-candidate list and domain thesaurus. The review of the previous user will be formalized into the preference keyword set of user.

4.4.2 Extraction

(a) Preprocess

HTML tags and stop words in the reviews snippet collection should be removed to avoid the quality of the keyword extraction getting affected in the next stage**.** Porter Stemmer algorithm is used to remove the commoner morphological and in flexional 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 of preferences**.** If the review contains a word which is also available in the domain thesaurus, the corresponding keyword should be extracted into the preference keyword set of the user.

4.4.3 Similarity Computation

The third phase of the proposed system is to identify the reviews of previous users who have similar tastes to the current user by finding neighborhoods of that user based on the similarity of their preferences. Before doing similarity computation, the reviews unrelated to the current user's preferences will be extracted by the intersection concept of set theory**.** If the result of intersection operation of the preference keyword sets of the current user and a previous user is an empty set, then the preference keyword set of the previous user will be filtered out. The Natural Language Processing is implemented to analyze the reviews of the previous user. The NLP Process comprises tokenizing a sentence or a word, parts of speech, tagging, the extraction operation of nouns and verbs, synonym retrieval of extracted keywords using WordNet dictionary**.** The Big Data manipulations from CSV are through our Own JAVA API that enforces developer friendly access.

5. Conclusion and Future Work

In the proposed system service recommendation through collaborative filtering method is used. Here keywords are used to indicate users' preferences, and a user-based Collaborative Filtering algorithm is adopted to generate appropriate recommendations. A keyword-candidate list and domain thesaurus are provided to help obtain users' preferences. The current 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. This method aims at presenting a personalized service recommendation list and recommending the most appropriate service(s) to the users. Moreover, to improve the scalability and efficiency of the “Big Data” environment, MapReduce framework is implemented. The experimental results demonstrate that this project significantly improves the accuracy and scalability of service recommender systems over existing approaches. As a future work the recommender application needs to be expanded across various domains. Along with that the application could be boosted with additional AI capabilities to provide its own rating for a service.

References

[1] J. Manyika, M. Chui, B. Brown, et al, "Big Data: The next frontier for innovation, competition, and productivity," 2011. C. Lynch, "Big Data: How do your data grow?" Nature, Vol. 455, No. 7209, pp. 28-29, 2008.
[2] F. Chang, J. Dean, S. Ghemawat, and W. C. Hsieh, "Bigtable: A distributed storage system for structured data," ACM Transactions on Computer Systems,Vol. 26, No. 2 (4) , 2008.
[3] W. Dou, X. Zhang, J. Liu, J. Chen, "HireSome-II: Towards Privacy- Aware Cross-Cloud Service Composition for Big Data Applications," IEEE Transactions on Parallel and Distributed Systems, 2013.
[4] 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.
[5] M. Bjelica, "Towards TV Recommender System Experiments with User Modeling," IEEE Transactions on Consumer Electronics, Vol.56, No.3, pp. 1763-1769, 2010.
[6] M. Alduan, F. Alvarez, J. Menendez, and O. Baez, "Recommender System for Sport Videos Based on User Audiovisual Consumption,"IEEE Transactions on Multimedia, Vol. 14, No.6, pp. 1546-1557, 2013.
[7] Y. Chen, A. Cheng and W. Hsu, "Travel Recommendation by Mining People Attributes and Travel Group Types From Community-Contributed Photos". IEEE Transactions on Multimedia, Vol. 25,No.6, pp. 1283-1295, 2012.
[8] Z. Zheng, X Wu, Y Zhang, M Lyu, and J Wang, "QoS Ranking Pre-diction for Cloud Services," IEEE Transactions on Parallel and Distributed Systems, Vol. 24, No. 6, pp. 1213-1222, 2013.
[9] W. Hill, L. Stead, M. Rosenstein, and G. Furnas, "Recommending and Evaluating Choices in a Virtual Community of Use," In CHI '95
[10] Proceedings of the SIGCHI Conference on Human Factors in Computing System, pp. 194-201, 1995.
[11] P. Resnick, N. Iakovou, M. Sushak, P. Bergstrom, and J. Riedl,
[12] "GroupLens: An Open Architecture for Collaborative Filtering of Netnews," In CSCW '94 Proceedings of the 1994 ACM conference on Computer supported cooperative work, pp. 175-186, 1994.
[13] R. Burke, "Hybrid Recommender Systems: Survey and Experiments," User Modeling and User-Adapted Interaction, Vol. 12, No.4, pp. 331-370, 2002.
[14] G. Adomavicius, and A. Tuzhilin, "Toward the Next Generation of Recommender Systems: A Survey of the Stateof-the-Art and Possible Extensions," IEEE Transactions on Knowledge and Data Engineering, Vol.17, No.6 pp. 734-749, 2005.
[15] D. Agrawal, S. Das, A. El Abbadi, "Big Data and cloud computing: new wine or just new bottles?" Proceedings of the VLDB Endowment, Vol. 3, No.1, pp. 1647-1648, 2010.
[16] J. Dean, and S. Ghemawat, "MapReduce: Simplified data processing on large clusters," Communications of the ACM, Vol. 51, No.1, pp. 107-113, 2005.
[17] G. DeCandia, D. Hastorun, M. Jampani, G. Kakulapati, A. Lakshman, A. Pilchin, S. Sivasubramanian, P. Vosshall, and W. Vogels,
[18] "Dynamo: Amazons highly available key-value store," In: Proceedings of the 21st ACM Symposium on Operating Systems Principles, pp. 205-220, 2007.
[19] M. Isard, M. Budiu, Y. Yu, A. Birrell, and D. Fetterly, "Dryad:Distributed data-parallel programs from sequential build-ing blocks," European Conference on Computer Systems, pp. 59-72, 2007.
[20] S. Ghemawat, H. Gobioff, and S. T. Leung, "The Google File Sys-tem," The 19th ACM Symposium on Operating Systems Principles, pp. 29-43, 2003.
[21] L. Zhang, "Editorial: Big Services Era: Global Trends of Cloud Computing and Big Data". IEEE Transactions on Services Computing, Vol. 5, No. 4, pp. 467-468, 2012.
[22] Z. Luo, Y. Li and J. Yin, "Location: a feature for service selection in the era of big data," 2013 IEEE 20th International Conference on Web Service, pp. 515-522, 2013.

Author Profile

M.Vigneesh received his undergraduate degree in Computer Science and Engineering in 2012 from Sri Sairam Engineering College. He received his post graduate degree in Cloud Computing in 2015 from SRM University. His areas of interest are Virtualization and Cloud Architecture and has hands on experience with Openstack. He has organized seminars on Cloud Technology and Virtualization on several colleges.

K.Nimala received her undergraduate degree in Electronics and Communication Engineering in 1999 from St.Joseph's College of Engineering and her post graduate degree in Computer Science and Engineering in 2000 from Sathyabama Engineering College. She was working as Assistant Professor(Sr.G) in Information Technology from 2000 to 2006. Later she worked on several IT companies from 2006 to 2012 and is now continuing her lecturer profile. Her research interests are Data Mining and Database Technology. Currently she works on papers like Personal Ontology based on User profile and Personalization -user adaptive website using data mining approach.