e-ISSN: 2349-9745
p-ISSN: 2393-8161

Scientific Journal Impact Factor (SJIF): 1.711

International Journal of Modern Trends in Engineering and Research

www.ijmter.com

Collaborative Based Clustering On Big Data Using HACE Theorem

Saravanan.N[1], Chinnadurai.S[2], DineshRaj.R[3], Suguna.G[4]
[1]*Department of ME-CSE, Srinivasan Engineering College*
[2]*Assistant Professor/CSE, Srinivasan Engineering College*
[3]*Department of ME-CSE, Srinivasan Engineering College*
[4]*Department of ME-CSE, Jayaram College of Engineering &Techonology*

Abstract— Currently, the large amount of data produced by many organizations is outpacing their storage ability. The management of such huge amount of data is quite expensive due to the requirements of high storage capacity and qualified personnel. So the Big Data is used to describe these types of massive volume of both structured and unstructured data that is so large that it's difficult to process using traditional database and software techniques. Big Data has the potential to help companies improve operations and make faster, more intelligent decisions. In many E-Commerce sites, Recommender Systems (RS), which provide personalized recommendation from among a large number of items, are recently introduced. Collaborative clustering is one of the most successful algorithms which provide recommendations using ratings of users on items. A CLUBCF (Clustering-Based Collaborative Filtering) approach is a service and its aims at gathering similar services in the same clusters to the recommend services collaborativelyThe HACE theorem is a data-driven model and it involves demand-driven aggregation of information sources, mining and analysis, user interest modeling and security and privacy considerations. At the filtering stage HACE theorem is to be applied for efficient process. At last, several important experiments are performed to verify the availability of the approach

Keywords— *Big Data, CLUBCF,HACE Theorem,Recommender System(RS)*

I. INTRODUCTION

Big Data has emerged as a widely recognized trend, attracting attentions from government, industry and academia. Big Data concerns large-volume, complex, growing data sets with multiple, autonomous sources. Big Data applications where data collection has grown tremendously and is beyond the ability of commonly used software tools to capture, manage, and process. It provide a new means to sense the public interests and generate feedback in real-time and are mostly appealing compared to generic media, such as radio or TV broadcasting.

Public picture sharing site, which received 1.8 million photos per day. On averaging, assuming the site of the each photo is 2MB, this requires 3.6TB storage every day. The most fundamental challenge for the Big Data applications is to explore the large volumes of data and extract useful information or knowledge for future actions.

In many situations, the knowledge extraction process has to be very efficient and close to real time because storing all observed data is nearly infeasible. Service users have nowadays encounter unprecedented difficulties in finding ideal ones from the overwhelming services.

Recommender Systems (RSs) are techniques and intelligent applications to assist users in a decision making process where they want to by weighted sum of description similarities and functionality similarities. Then, services are merged into clusters according to their characteristic similarities.

Clustering are techniques that can reduce the data sizes by a large factor by grouping similar services together. Therefore, the main objective is Clustering-based Collaborative Filtering approach (ClubCF) and HACE (Heterogeneous Autonomous Complex Evolving) theorems are used. ClubCF consists of two stages: clustering and collaborative filtering. Clustering is a preprocessing step to separate Big Data into manageable parts.

A cluster contains some similar services just like a club contains some like-minded users. This is another reason besides abbreviation for ClubCF. Since the number of services in a cluster is much less than the total number of services, the computation time of CF (Collaborative Filtering) algorithm can be reduced significantly.

Big Data starts with large-volume, heterogeneous, autonomous sources to explore complex and evolving relationship among data. These characteristics make it an extreme challenge for discovering useful knowledge from the Big Data. Exploring these data in this scenario is equivalent to aggregating heterogeneous information from different sources

II. RELATED WORK

Big Data applications deals with large number of data sets and it is usually measures in peta bytes. The concept of BigTable is used to store these data. A BigTable is a sparse, distributed, persistent multi dimensional stored map. The map is indexed by a row key, column key and a timestamp. First step of the work is to design BigTable for storage requirement of CluBCF. In the second step, CluBCF approach is designed and its characteristic similarities between services and are computed by weighted sum of description similarities and functionality similarities.
Then the services are merged into clusters according to their characteristic similarities. Further, an item-based CF algorithm is applied within the cluster that the target service belongs to. The final step several experiments are conducted on a real dataset extracted from amazon (https://aws.amazon.com/datasets).

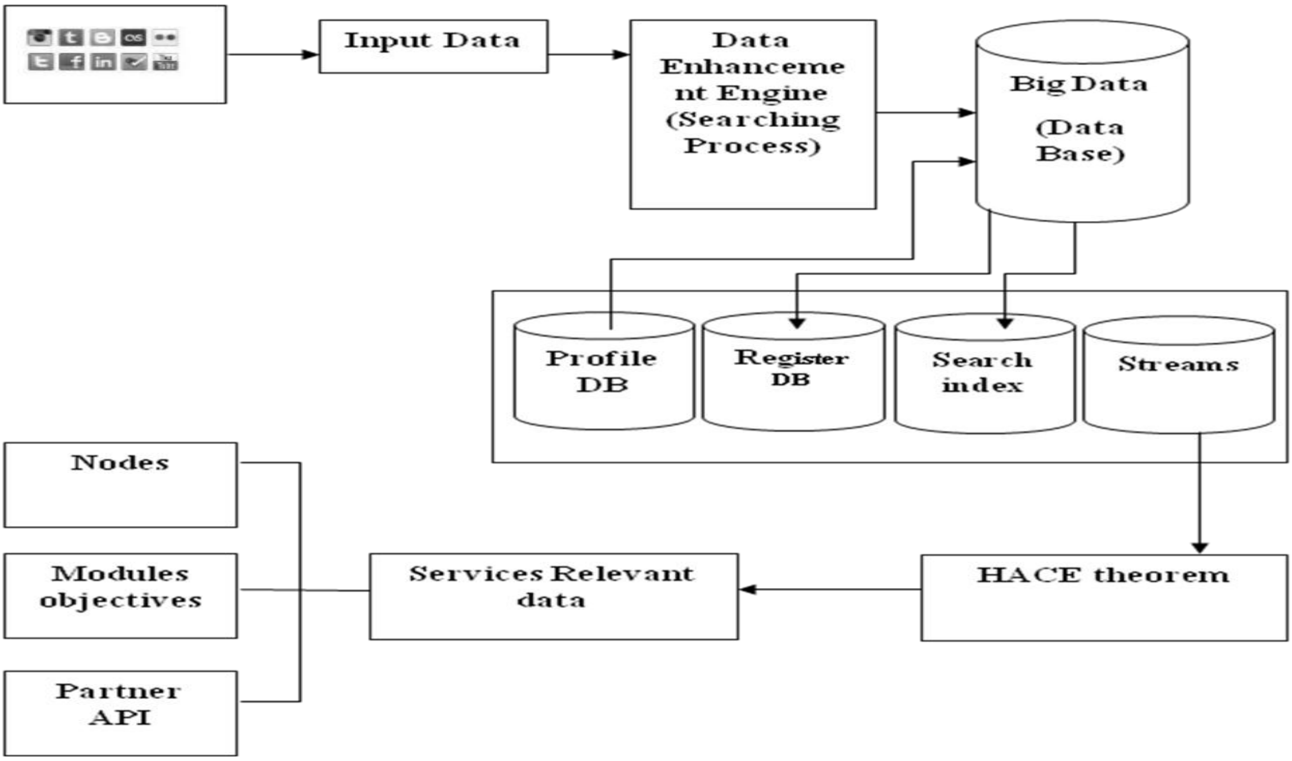

Figure 1 system architecture

To measure the similarity between Web services, investigated the metadata from the WSDL (Web Service Description Language) files and defined a Web service as = N,M,D,O , where N is the name that specifies a

Web service, M is the set of messages exchanged by the operation invocation, D is the set of data types, and O is the set of operations provided by the Web service. From the definition, three types of metadata from WSDL can be identified for similarity matching: the plain textual descriptions, the operation that captures the purposed functionality and the data type relate to the semantic meanings. For evaluating reputation of service, Li et al. defined a Web service as $WS(id,d,t,sg,rs,dor)$ where id is its identity, d is its text description, t is its classification, sg denotes the level of its transaction volume, rs is its review set, and dor is its reputation degree. In the SOA Solution Stack (S3) proposed by IBM, a service is defined as an abstract specification of one or more business-aligned IT functions. This specification provides consumers with sufficient information to be able to invoke the business functions exposed by a service provider.

Although the definitions of service are distinct and application-specific, they have common elements which mainly include service descriptions and service functionalities. In addition, rating is an important user activity that reflects their opinions on services. Especially in application of service recommendation, service rating is an important element. As more and more services are emerging on the Internet, such huge volume of service-relevant elements are generated and distributed across the network, which cannot be effectively accessed by traditional database management system. To address this problem, Bigtable is used to store services in this paper. Bigtable is a distributed storage system of Google for managing structured data that is designed to scale to a very large size across thousands of commodity servers.

A Bigtable is a sparse, distributed, persistent multi-dimensional sorted map. The map is indexed by a row key, column key, and a timestamp; each value in the map is a un interpreted array of bytes. Column keys are grouped into sets called column families, which form the basic unit of access control. A column key is named using the following syntax: family, qualifier, where, family‟ refers to column family and “qualifier‟ refers to column key. Each cell in a Bigtable can contain multiple versions of the same data which are indexed by timestamp. Different versions of a cell are stored in decreasing timestamp. With this assumption, a ClubCF approach for Big Data application is presented, which aims at recommending services from overwhelming candidates within an acceptable time. Technically, ClubCF focuses on two interdependable stages, i.e., clustering stage and collaborative filtering stage. In the first stage, services are clustered according to their characteristic similarities. In the second stage, a collaborative filtering algorithm is applied within a cluster that a target service it belongs.

Different developers may use different-form words to describe similar services. Using these words directly may influence the measurement of description similarity.

Therefore, description words should be uniformed before further usage. In fact, morphological similar words are clubbed together under the assumption that they are also semantically similar. For example, “map”, “maps” and “mapping” are forms of the equivalent lexeme, with “map‟ as the morphological root form. To transform variant word forms to their common root called stem, various kinds of stemming algorithms, such as Lovins stemmer, Dawson Stemmer, Paice/Husk Stemmer, and Porter Stemmer, have been proposed. Among them, Porter Stemmer is one of the most widely used stemming algorithms

HACE Theorem

- **H**eterogeneous, **A**utonomous, **C**omplex, **E**volving
- Big Data starts with large volume, heterogeneous, autonomous sources with distributed and decentralized control, and seeks to explore complex and evolving relationships among data
- These are characteristics of Big Data
- This is theorem to model Big Data characteristics
- Huge Data with heterogeneous and diverse dimensionality
- represent huge volume of data
- Autonomous sources with distributed and decentralized control

- main characteristics of Big Data
- Complex and evolving relationships

Compute Description Similarity and Functionality Similarity

Description similarity and functionality similarity are both computed by Jaccard similarity coefficient (JSC) which is a statistical measure of similarity between samples sets. For two sets, JSC is defined as the cardinality of their intersection divided by the cardinality of their union.

Compute Characteristic Similarity

Characteristic similarity between st and sj is computed by weighted sum of description similarity and functionality similarity.

Cluster Services

Clustering is a critical step in our approach. Clustering methods partition a set of objects into clusters such that objects in the same cluster are more similar to each other than objects in different clusters according to some defined criteria. Generally, cluster analysis algorithms have been utilized where the huge data are stored. Clustering algorithms can be either hierarchical or partitional. Some standard partitional approaches (e.g., *K*-means) suffer from several limitations: 1) results depend strongly on the choice of number of clusters *K*, and the correct value of *K* is initially unknown; 2) cluster size is not monitored during execution of the *K*-means algorithm, some clusters may become empty ("collapse"), and this will cause premature termination of the algorithm; 3) algorithms converge to a local minimum. Hierarchical clustering methods can be further classified into agglomerative or divisive, depending on whether the clustering hierarchy is formed in a bottom-up or top-down fashion.

Many current state-of-the-art clustering systems exploit agglomerative hierarchical clustering (AHC) as their clustering strategy, due to its simple processing structure and acceptable level of performance.

Select Neighbors

Based on the enhanced rating similarities between services, the neighbors of a target service st

$N\ st = sj\ R\_sim'\ st,sj > \gamma\ ,st \neq sj$

Here, R_sim' , is the enhanced rating similarity between service st and sj computed by γ is a rating similarity threshold. The bigger value of γ is, the chosen number of neighbors wills relatively less but they may be more similar to the target service, thus the coverage of collaborative filtering will decrease but the accuracy may increase. On the contrary, the smaller value of γ is, the more neighbors are chosen but some of them may be only slightly similar to the target service, thus the coverage of collaborative filtering will increase but the accuracy would decrease. Therefore, a suitable γ should be set for the tradeoff between accuracy and coverage. While γ is assigned, sj will be selected as a neighbor of st and put into the neighbor set $N\ st$ if $R\_sim'\ st, > \gamma$.

Compute Predicted Rating

For an active user ua for whom predictions are being made, whether a target service st is worth recommending depends on its predicted rating. If $(st) \neq \Phi$, similar to the computation formula proposed by the predicted rating $P(ua,st)$ in an item-based CF is computed as follow:

$Pua,st = rst + rua,sj - rsj \times R\_sim'\ st,sj\ sj \in N(st)\ R\_sim'\ st,sj\ sj \in N(st)$

Here, $rst$is the average rating of st, (st) is the neighbor set of st, $sj \in (st)$ denotes sj is a neighbor of the target service st, rua,sj is the rating that an active user ua gave to sj, rsj is the average rating of sj, and $R\_sim'\ st,sj$ is the enhanced rating

III. SCOPE OF THE PAPER

Different developers may use different-form words to describe similar services. Using these words directly may influence the measurement of description similarity. Therefore, description words should be uniformed before further usage. In fact, morphological similar words are clubbed together under the assumption that they are also semantically similar. For example, "map", "maps" and "mapping" are forms of the equivalent lexeme, with "map‟ as the morphological root form. To transform variant word forms to their common root called stem, various kinds of stemming algorithms, such as Lovins stemmer, Dawson Stemmer, Paice/Husk Stemmer, and Porter Stemmer, have been proposed. Among them, Porter Stemmer is one of the most widely used stemming algorithms. It applies cascaded rewrite rules that can be run very quickly and do not require the use of a lexicon

Data Gathering

Big Data is a popular term used to describe the exponential growth and availability of data, both structured and unstructured. And Big Data may be as important to business, society and the Internet. Big Data is an all-encompassing term for any collection of data sets so large and complex that it becomes difficult to process using traditional data processing applications. Big Data usually includes data sets with sizes beyond the ability of commonly used software tools to capture, curate, manage, and process data within a tolerable elapsed time..

Data Pre-Processing

The query submitted by the user contains parts of speech and special characters which are not required for analysis as they do not truly reflect the relevance of a search result. If this query is used for analysis, it may give inconsistent and inaccurate results. Therefore, the user query will be pre-processed to identify the root words.

Query Based Clustering

Agglomerative Hierarchical Clustering to cluster the pseudo – documents. Name of the algorithm refers to its way of working, as it creates hierarchical results in an "agglomerative" or "bottom-up" way, i.e. by merging smaller groups into larger ones. Algorithm takes as input a matrix of pair wise similarities between objects. In case of documents this matrix is created by calculating all pair wise similarities between documents using cosine similarity. It returns a binary tree of clusters, called dendrogram. Clustering is a preprocessing step to separate Big Data into manageable parts.

Collaborative Filtering

Collaborative filtering (CF) is a technique. Collaborative filtering is a method of making automatic predictions. All services are stored in a table which is called service table. The corresponding elements will be drawn from service table during the process of CF.

Exhaustive Search

The search proceeds by generating and testing each node that is reachable from a parent node before it expands any of these children. Exhaustive systematic search is referred to a breadth-first search. The system retrieves the stored records into memory. And find that the fastest method is loading all data (about 320 MB) at once with one SQL query, instead of fetching one by one

Rating

The system upgrades the search results which include the emphasized term or sentence. The system rating the search results according to the user intention and shows the rating results to the user.

IV. CONCLSION

Driven by real-world applications and key industrial stakeholders and initialized by national funding agencies, managing and mining Big Data have shown to be a challenging yet compelling task. While the term Big Data literally concerns about data volumes, a ClubCF approach using HACE theorem for Big Data applications and relevant services. Before applying CF technique, services are merged into some clusters via an AHC algorithm. Then the rating similarities between services within the same cluster are computed. As the number of services in a cluster is much less than that of in the whole system, ClubCF costs less online computation time. Moreover, as the ratings of services in the same cluster are more relevant with each other than with the ones in other clusters, prediction based on the ratings of the services in the same cluster will be more accurate than based on the ratings of all similar or dissimilar services in all clusters. These two advantageous of ClubCF and HACE have been verified by experiments on real-world data set. To explore Big Data, challenges are analyzed, model and system level. To support Big Data mining, high-performance computing platforms are required, which impose systematic designs to unleash the full power of the Big Data.

REFERENCES

[1] Ahmed. R and Karypis. G. (2012), 'Algorithms for Mining the Evolution of Conserved Relational States in Dynamic Networks', Knowledge and Information Systems, vol. 33, no. 3, pp. 603-630.
[2] Bellogín. A., Cantador. I and Díez .F. (2013), 'An empirical comparison of social, collaborative filtering, and hybrid recommenders', ACM Trans. on Intelligent Systems and Technology, vol. 4, no. 1, pp. 1-37.
[3] Catherine. M. R and Edwin. E. B. (2013), 'A Survey on Recent Trends in Cloud Computing and its Application for Multimedia', International Journal of Advanced Research in Computer.
[4] Chang .F., Dean .J and Mawat .S. (2008), 'Bigtable: A distributed storage system for structured data', ACM Trans. on Computer Systems, vol. 26, no. 2, pp. 1-39.
[5] Chen. R., Sivakumar. K and Kargupta. H. (2004), 'Collective Mining of Bayesian Networks from Distributed Heterogeneous Data', Knowledge and Information Systems, vol. 6, no. 2, pp. 164-187.
[6] Elmeleegy .H., Ivan. A and Akkiraju. R. (2008), 'Mashup advisor: A recommendation tool for mashup development,' in Proc. of IEEE Int'l Conf. on Web Services, pp. 337-344.
[7] Greenshpan. O., Milo. T and Polyzotis. N. (2009), 'Autocompletion for mashups' in Proc. of the VLDB Endowment, vol. 2, no. 1, pp. 538-549.
[8] Gupta. V and Lehal. G. S. (2013), 'A Survey of Common Stemming Techniques and Existing Stemmers for Indian Languages', Journal of Emerging Technologies in Web Intelligence, vol. 5, no. 2, pp. 157-161.
[9] Herlocker. J. L., Konstan. J. A and Terveen. L. G. (2004), 'Evaluating collaborative filtering recommender system' ACM Trans. on Information Systems, vol. 22, no. 1, pp. 5-53
[10] Li. H. H., Du. X. Y, and Tian. X. (2009), 'A review-based reputation evaluation approach for Web services,' Journal of Computer science and technology, vol. 24, no. 5, pp. 893-900
[11] Liu. X., Hui. Y and Sun. W. (2007), 'Towards service composition based on mashup', in Proc. of IEEE Congress on Services, pp. 332-339.
[12] Liu. X., Huang. X and Mei. H. (2009), 'Discovering homogeneous web service community in the user-centric web environment', IEEE Trans. on Services Computing, vol. 2, no. 2, pp. 167-181.
[13] Liu. Z., Li. P and Zheng. Y. (2009), 'Clustering to find exemplar terms for keyphrase extraction' in Proc. 2009 Conf. on Empirical Methods in Natural Language Processing, pp. 257-266.
[14] Mojgan Ghanavati., Mohamad Reza Gholamian., Behrouz Minaei and Mehran Davoudi. (2011), 'An Efficient Cost Function For Imperialist Competitive Algorithm To Find Best Clusters'.
[15] Sandeep. R. S., Vinay, C and Hemant. S. M. (2013), 'Strength and Accuracy Analysis of Affix Removal Stemming Algorithms' International Journal of Computer Science and Information Technologies, vol. 4, no. 2, pp. 265-269.
[16] Singaravelu.S., Sherin .A and Savitha. S. (2013), 'Agglomerative Fuzzy K-Means Clustering Algorithm'.
[17] SongJie Gong. (2010), 'A Collaborative Filtering Recommendation Algorithm Based on User Clustering and Item Clustering'.
[18] A. Yamashita et al, (2003), 'Adaptive Fusion Method For User-Based And Item-Based Collaborative Filtering
vol. 15, no. 2, pp. 353-367.
[19] Zielinnski,(2010), Adaptive SOA Solution Stack.
.