ISSN 2348-1196 (print)
International Journal of Computer Science and Information Technology Research ISSN 2348-120X (online)
Vol. 3, Issue 2, pp: (95-101), Month: April - June 2015, Available at: www.researchpublish.com

Service Oriented Analysis of Big Data Using CLUBCF

[1]R.Barjana, [2]C.Suganthi, [3]A.Deebiga, [4]G.Ashwini, [5]G.Abirami,

[1,2,3,4] Students, [5]HOD, Dept of Information Technology, V.S.B Engineering College, karur. India

Abstract: **Big Data is an all-inclusive term for any gathering of information sets so huge and complex that it gets to be hard to process utilizing conventional information handling applications. The difficulties incorporate investigation, capture, search, sharing, storage, exchange, visualization, and security infringement. The pattern to bigger information sets is because of the extra data logical from examination of a solitary huge set of related information, as contrasted with independent more diminutive sets with the same aggregate sum of information, permitting associations to be found to "spot business patterns, avoid maladies, battle wrongdoing etc. In this paper we proposed a semantic analysis and trust analysis because in semantic analysis text mining approach is utilized for survey the item and discover the related services. Trust Analysis is utilized for discover whether is trustable or non-trustable.**

Keywords: **big data application, cluster, Text mining, Semantic analysis, Trust Analysis**

I. INTRODUCTION

Big Data has risen as a broadly perceived pattern, drawing in considerations from government, industry and the scholarly world. As a rule, Big Data concerns extensive volume, perplexing, developing information sets with numerous, self-ruling sources. Big Data applications where information gathering has developed colossally and is past the capacity of ordinarily utilized programming instruments to catch, oversee, and handle inside a "tolerable passed time" is on the ascent. The most crucial test for the Big Data applications is to investigate the substantial volumes of information and concentrate helpful data or learning for future activities.

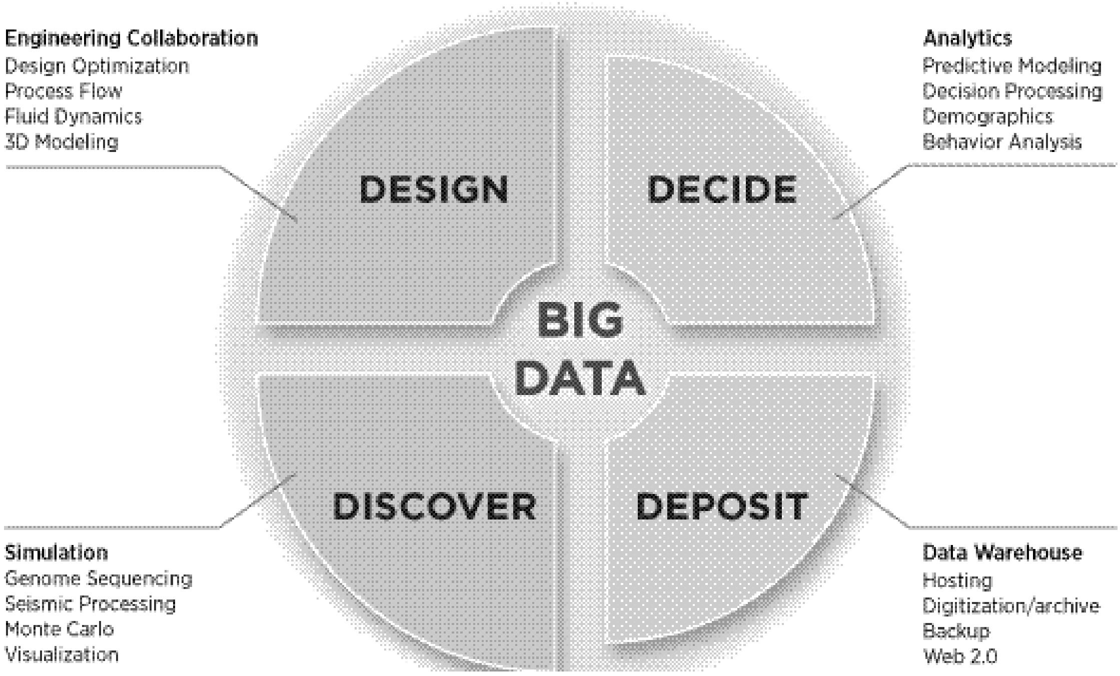


Figure 1: Big Data

Text mining, also referred to as text data mining, roughly equivalent to text analytics, refers to the process of deriving high-quality information from text. High-quality information is typically derived through the devising of patterns and trends First, data locality may result in a waste of resources. For example, most computation resource of a server with less popular data may stay at rest. The low resource efficacy through means such as statistical pattern learning. Text mining usually involves the process of structuring the input text (usually parsing, along with the addition of some derived linguistic features and the removal of others, and subsequent insertion into a database), deriving patterns within the structured data, and finally evaluation and interpretation of the output. 'High quality' in text mining usually refers to some combination of relevance, novelty, and interestingness.

Typical text mining tasks include text categorization, text clustering, concept/entity extraction, production of granular taxonomies, sentiment analysis, summarization document summarization, and entity relation modeling. Text analysis involves information retrieval, lexical analysis to study word frequency distributions, pattern recognition, tagging/annotation, information extraction, data mining techniques including link and association analysis, visualization, and predictive analytics. The overarching goal is, essentially, to turn text into data for analysis, via application of natural language processing (NLP) and analytical methods.

Text mining is the analysis of data contained in natural language text. Text mining works by transposing words and phrases in unstructured data into numerical values which can then be linked with structured data in a database and analyzed with traditional data mining techniques. Compared with the kind of data stored in databases, text is unstructured, amorphous, and difficult to deal with algorithmically. Nevertheless, in modern culture, text is the most common vehicle for the formal exchange of information. The field of text mining usually deals with texts whose function is the communication of factual information or opinions, and the motivation for trying to extract information from such text automatically is compelling even if success is only partial. Success is only partial.

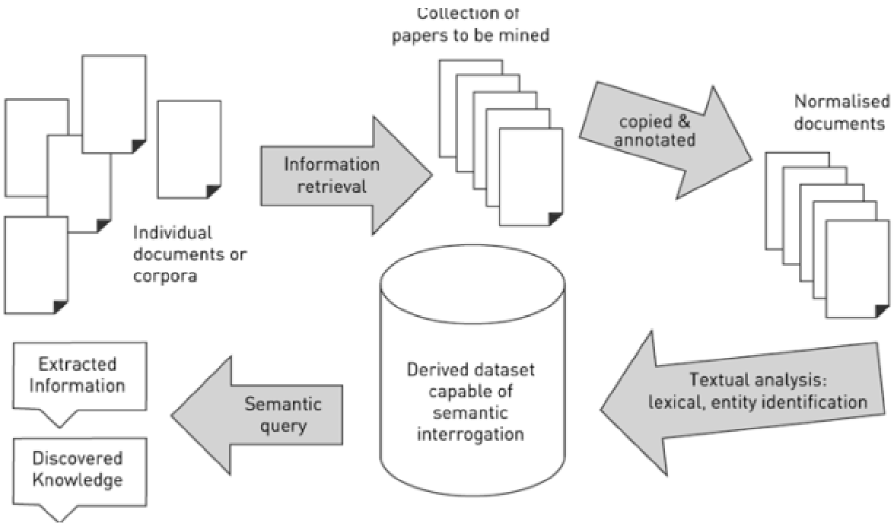

Collection of papers to be mined
Information retrieval
copied & annotated
Normalised documents
Individual documents or corpora
Derived dataset capable of semantic interrogation
Textual analysis: lexical, entity identification
Extracted Information
Discovered Knowledge
Semantic query

Figure 2: Text Mining Process

Cluster analysis or clustering is the task of grouping a set of objects in such a way that objects in the same group (called a cluster) are more similar (in some sense or another) to each other than to those in other groups (clusters). It is a main task of exploratory data mining, and a common technique for statistical data analysis, used in many fields, including machine learning.

Cluster analysis itself is not one specific algorithm, but the general task to be solved. It can be achieved by various algorithms that differ significantly in their notion of what constitutes a cluster and how to efficiently find them. Popular notions of clusters include groups with small distances among the cluster members, dense areas of the data space, intervals or particular statistical distributions. Clustering can therefore be formulated as a multi-objective optimization problem. The appropriate clustering algorithm and parameter settings (including values such as the distance function to use, a density threshold or the number of expected clusters) depend on the individual data set and intended use of the results. Cluster analysis as such is not an automatic task, but an iterative process of knowledge discovery or interactive multi-objective optimization that involves trial and failure. It will often be necessary to modify data preprocessing and model parameters until the result achieves the desired properties.

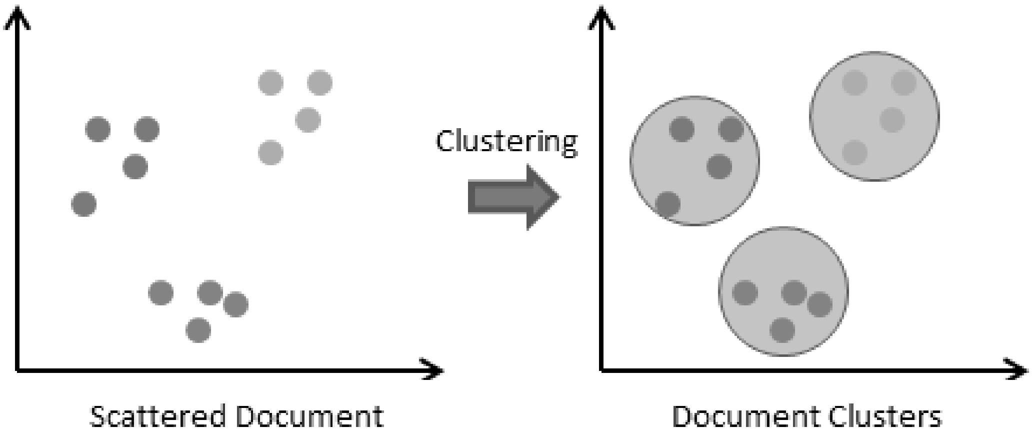


Figure 3: Cluster Formation

2. RELATED WORK

In [1] X. Wu, X. Zhu, G. Q. Wu, et al., Big Data are now rapidly expanding in all science and engineering domains, including physical, biological and biomedical sciences. This paper presents a HACE theorem that characterizes the features of the Big Data revolution, and proposes a Big Data processing model, from the data mining perspective. This data-driven model involves demand-driven aggregation of information sources, mining and analysis, user interest modeling, and security and privacy considerations. We analyze the challenging issues in the data-driven model and also in the Big Data revolution.

In [2] Wei Fan, Albert Bifet., Big Data mining is the capability of extracting useful information from these large datasets or streams of data, that due to its volume, variability, and velocity, it was not possible before to do it. The Big Data challenge is becoming one of the most exciting opportunities for the next years. We present in this issue, a broad overview of the topic, its current status, controversy, and forecast to the future. We introduce four articles, written by influential scientists in the field, covering the most interesting and state-of-the-art topics on Big Data mining.

In [3] X. Li, and T. Murata., In this paper, we present a hybrid recommendation approach for discovering potential preferences of individual users. The proposed approach provides a flexible solution that incorporates multidimensional clustering into a collaborative filtering recommendation model to provide a quality recommendation. This facilitates to obtain user clusters which have diverse preference from multi-view for improving effectiveness and diversity of recommendation. The presented algorithm works in three phases: data preprocessing and multidimensional clustering, choosing the appropriate clusters and recommending for the target user. The performance of proposed approach is evaluated using a public movie dataset and compared with two representative recommendation algorithms. The empirical results demonstrate that our proposed approach is likely to trade-off on increasing the diversity of recommendations while maintaining the accuracy of recommendations.

3. CLUB CF

In Existing system Clustering-based Collaborative Filtering approach (ClubCF) is used which aims at recruiting similar services in the same clusters to recommend services collaboratively. Collaborative filtering (CF) is a technique used by some recommender systems. Collaborative filtering has two senses, a narrow one and a more general one. In general, collaborative filtering is the process of filtering for information or patterns using techniques involving collaboration among multiple agents, viewpoints, data sources, etc. Applications of collaborative filtering typically involve very large data sets. Collaborative filtering methods have been applied to many different kinds of data including: sensing and monitoring data, such as in mineral exploration, environmental sensing over large areas or multiple sensors; financial data, such as financial service institutions that integrate many financial sources; or in electronic commerce and web applications where the focus is on user data,

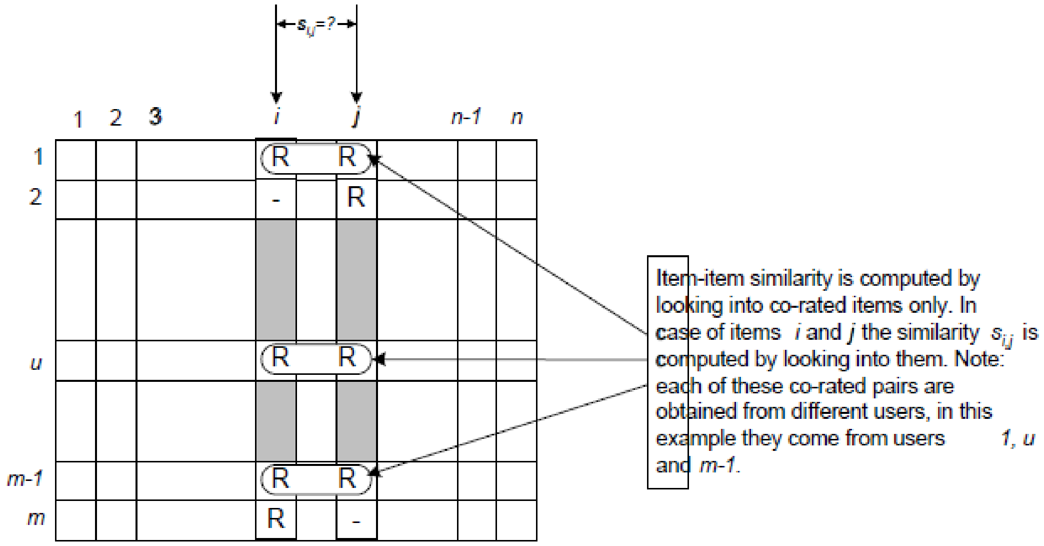
etc. The remainder of this discussion focuses on collaborative filtering for user data, although some of the methods and approaches may apply to the other major applications as well.

In the newer, narrower sense, collaborative filtering is a method of making automatic predictions (filtering) about the interests of a user by collecting preferences or taste information from many users (collaborating). The underlying assumption of the collaborative filtering approach is that if a person *A* has the same opinion as a person *B* on an issue, A is more likely to have B's opinion on a different issue *x* than to have the opinion on x of a person chosen randomly. For example, a collaborative filtering recommendation system for television tastes could make predictions about which television show a user should like given a partial list of that user's tastes (likes or dislikes). Note that these predictions are specific to the user, but use information gleaned from many users. This differs from the simpler approach of giving an average (non-specific) score for each item of interest, for example based on its number of votes.

3.1 Item Based Collaborative Filtering Algorithm:

Item-based approaches apply the same idea, but use similarity between items instead of users. To determine the most-similar match for a given item, the algorithm builds a similar-items table by finding items that customers tend to purchase together. We could build a product-to-product matrix by iterating through all item pairs and computing a similarity metric for each pair. However, many product pairs have no common customers, and thus the approach is inefficient in terms of processing time and memory usage. The following iterative algorithm provides a better approach by calculating the similarity between a single product and all related products:

It's possible to compute the similarity between two items in various ways, but a common method is to use the cosine measure we described earlier, in which each vector corresponds to an item rather than a customer, and the vector's *M* dimensions correspond to customers who have purchased that item. This offline computation of the similar-items table is extremely time intensive, with *O*(*N*2*M*) as worst case. In practice, however, it's closer to *O*(*NM*), as most customers have very few purchases. Sampling customers who purchase best-selling titles reduces runtime even further, with little reduction in quality. Given a similar-items table, the algorithm finds items similar to each of the user's purchases and ratings, aggregates those items, and then recommends the most popular or correlated items. This computation is very quick, depending only on the number of items the user purchased or rated.



Item-item similarity is computed by looking into co-rated items only. In case of items *i* and *j* the similarity $s_{i,j}$ is computed by looking into them. Note: each of these co-rated pairs are obtained from different users, in this example they come from users *1, u* and *m-1*.

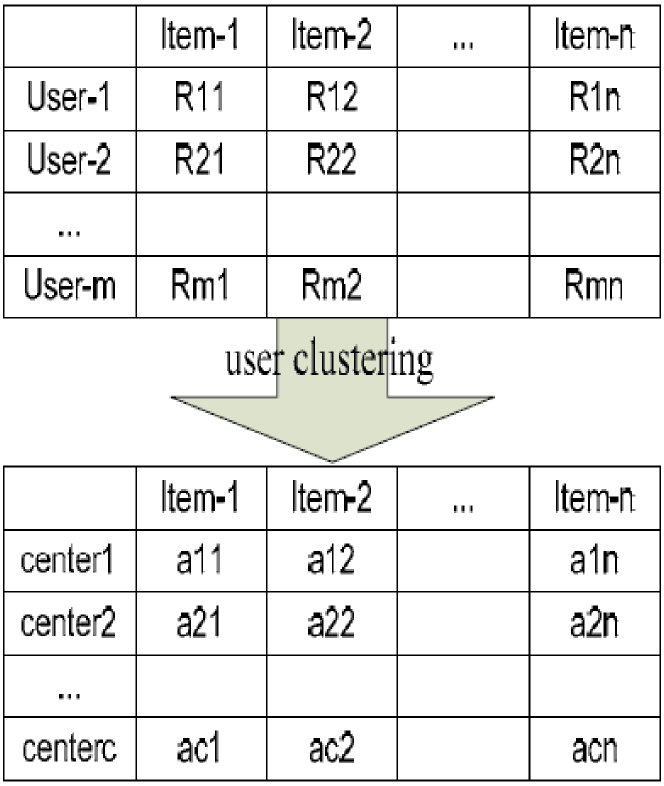
Figure 4: CF Algorithm Structure

Item clustering techniques work by identifying groups of items who appear to have similar ratings. Once the clusters are created, predictions for a target item can be made by averaging the opinions of the other items in that cluster. Some clustering techniques represent each item with partial participation in several clusters. The prediction is then an average across the clusters, weighted by degree of participation. Once the item clustering is complete, however, performance can be very good, since the size of the group that must be analyzed is much smaller.

The idea is to divide the items of a collaborative filtering system using item clustering algorithm and use the divide as neighborhoods. The clustering algorithm may generate fixed sized partitions, or based on some similarity threshold it may generate a requested number of partitions of varying size.

3.2 User Based collaborative filtering:

User clustering techniques work by identifying groups of users who appear to have similar ratings. Once the clusters are created, predictions for a target user can be made by averaging the opinions of the other users in that cluster. Some clustering techniques represent each user with partial participation in several clusters. The prediction is then an average across the clusters, weighted by degree of participation. Once the user clustering is complete, however, performance can be very good, since the size of the group that must be analyzed is much smaller. The idea is to divide the users of a collaborative filtering system using user clustering algorithm and use the divide as neighborhoods. The clustering algorithm may generate fixed sized partitions, or based on some similarity threshold it may generate a requested number of partitions of varying size.



	Item-1	Item-2	...	Item-n
User-1	R11	R12		R1n
User-2	R21	R22		R2n
...				
User-m	Rm1	Rm2		Rmn

	Item-1	Item-2	...	Item-n
center1	a11	a12		a1n
center2	a21	a22		a2n
...				
centerc	ac1	ac2		acn

Figure 5: User Based Collaborative Filtering

Where Rij is the rating of the user i to the item i, aij the average rating of the user center i to the item i, m is the number of all users, n is the number of all items, and c is the number of user centers.

4. SEMANTIC ANALYSIS

SA (Semantic Analysis) is a fully automatic mathematical/statistical technique for extracting and inferring relations of expected contextual usage of words in passages of discourse. It is not a traditional natural language processing or artificial intelligence program; it uses no humanly constructed dictionaries, knowledge bases, semantic networks, grammars, syntactic parsers, or morphologies, or the like, and takes as its input only raw text parsed into words defined as unique character strings and separated into meaningful passages or samples such as sentences or paragraphs.

The first step is to represent the text as a matrix in which each row stands for a unique word and each column stands for a text passage or other context. Each cell contains the frequency with which the word of its row appears in the passage denoted by its column. Next, the cell entries are subjected to a preliminary transformation, whose details we will describe later, in which each cell frequency is weighted by a function that expresses both the word's importance in the particular passage and the degree to which the word type carries information in the domain of discourse in general.

Next, SA applies singular value decomposition (SVD) to the matrix. This is a form of factor analysis, or more properly the mathematical generalization of which factor analysis is a special case. In SVD, a rectangular matrix is decomposed into the product of three other matrices. One component matrix describes the original row entities as vectors of derived values, another describes the original column entities in the same way, and the third is a diagonal matrix containing scaling values such that when the three components are matrix-multiplied, the original matrix is reconstructed. There is a mathematical proof that any matrix can be so decomposed perfectly, using no more factors than the smallest dimension of the original matrix. When fewer than the necessary number of factors are used, the reconstructed matrix is a least-squares best fit. One can reduce the dimensionality of the solution simply by deleting coefficients in the diagonal matrix, ordinarily starting with the smallest.

4.1 Natural Language Processing:

It is one of the oldest and most difficult problems in the field of artificial intelligence. It is the analysis of human language so that computers can understand natural languages as humans do. Although this goal is still some way off, NLP can perform some types of analysis with a high degree of success. Shallow parsers identify only the main grammatical elements in a sentence, such as noun phrases and verb phrases, whereas deep parsers generate a complete representation of the grammatical structure of a sentence. The role of NLP in text mining is to provide the systems in the information extraction phase (see below) with linguistic data that they need to perform their task. Often this is done by annotating documents with information like sentence boundaries, part-of-speech tags, parsing results, which can then be read by the information extraction tools.

In our project text mining is used for find out the user reviews and cluster the data clustering is act as a major role in our project because cluster is grouping the data and next time the user searching the same keyword it shows the related services to the user.

4.2 Text Mining Process:

Text mining involves the application of techniques from areas such as information retrieval, natural language processing, information extraction and data mining. These various stages of a text-mining process can be combined together into a single workflow**.**

Information Retrieval (IR) systems identify the documents in a collection which match a user’s query. The most well known IR systems are search engines such as Google, which identify those documents on the World Wide Web that are relevant to a set of given words. IR systems are often used in libraries, where the documents are typically not the books themselves but digital records containing information about the books. IR systems allow us to narrow down the set of documents that are relevant to a particular problem. As text mining involves applying very computationally-intensive algorithms to large document collections, IR can speed up the analysis considerably by reducing the number of documents for analysis. For example, if we are interested in mining information only about protein interactions, we might restrict our analysis to documents that contain the name of a protein, or some form of the verb ‘to interact’ or one of its synonyms.

4.3 Trust Analysis:

The trust-based clustering algorithms are classified into two types of algorithms: pure and hybrid. Regarding the first one, algorithms pursue two main purposes: first they aim to improve the network security by electing trustworthy services as CHs, second they try to decrease the overheads of trust management systems by combing trust related operations with various phases of clustering algorithms.

Trust Analysis is the important part of this project because finally it will analysis each and every web services as whether trustable or Non-trustable. Before we receive the related web services trust analysis will analysis that web services after only it shows, So users get the trustable services.

5. CONCLUSION AND FUTURE WORK

In proposed semantic analysis and trust analysis. Clients gives the evaluation and remarks for services our semantic analysis is group the comments(text) and it gathered them, if an alternate time when the client seek the same keyword our semantic analysis demonstrates the related web services for that keyword.

In future work we discover the cluster outlier motivation behind this is, when at some point clustering miss some service while gathering this outlier figure out the missing services and cluster them into the group, so client won't miss any services.

REFERENCES

[1] X. Wu, X. Zhu, G. Q. Wu, et al., "Data mining with big data," IEEE Trans. on Knowledge and Data Engineering, vol. 26, no. 1, pp. 97-107, January 2014.

[2] Wei Fan, Albert Bifet., "Mining Big Data: Current Status, and Forecast to the Future" Volume 14, Issue 2

[3] X. Li, and T. Murata."Using Multidimensional Clustering Based Collaborative Filtering Approach Improving Recommendation Diversity," in Proc. 2012 IEEE/WIC/ACM Int'l Joint Conf. on Web Intelligence and Intelligent Agent Technology, pp. 169-174, December 2012.

[4] F. Chang, J. Dean, S. mawat, et al., "Bigtable: A distributed storage system for structured data," ACM Trans. on Computer Systems, vol. 26, no. 2, pp. 1-39, June 2008.

[5] M. J. Li, M. K. Ng, Y. M. Cheung, et al. "Agglomerative fuzzy k-means clustering algorithm with selection of number of clusters," IEEE Trans. on Knowledge and Data Engineering, vol. 20, no. 11, pp. 1519-1534, November 2008.

[6] Z. Zhou, M. Sellami, W. Gaaloul, et al., "Data Providing Services Clustering and Management for Facilitating Service Discovery and Replacement," IEEE Trans. on Automation Science and Engineering, vol. 10, no. 4, pp. 1-16, October 2013.

[7] R. D. Simon, X. Tengke, and W. Shengrui, "Combining collaborative filtering and clustering for implicit recommender system," in Proc. 2013 IEEE 27th Int'l Conf. on. Advanced Information Networking and Applications, pp. 748-755, March 2013.

[8] Z. Zheng, H. Ma, M. R. Lyu, et al., "QoS-aware Web service recommendation by collaborative filtering," IEEE Trans. on Services Computing, vol. 4, no. 2, pp. 140-152, February 2011.

[9] Z. Liu, P. Li, Y. Zheng, et al., "Clustering to find exemplar terms for key phrase extraction," in Proc. 2009 Conf. on Empirical Methods in Natural Language Processing, pp. 257-266, May 2009.

[10] H. H. Li, X. Y. Du, and X. Tian, "A review-based reputation evaluation approach for Web services," Journal of Computer science and technology, vol. 24, no. 5, pp. 893