A Survey on Web Services Recommendation Model for Big data Environment

Sandeep Veerwani[1], Dr. M. K. Rawat[2]
[1, 2] Department of Computer Science & Engineering
[1, 2] LNCT, Indore

Abstract- *The recommendation is a correlation among the user needs and the available entities. Now in this world of technology a number of services providers are increasing continuously and offers various services. Sometimes for the similar services a number of providers are available who offers more attractive proposals. Thus selection among a number of services by the user is a complicated task, and the user is confused "which service is more relevant and suitable for satisfying user's need". The proposed work is intended to provide the method by which the optimum services are going to be suggested. The proposed technique includes the modelling of collaborative filtering to estimate the user requirement and produces the more relevant services for the end user. The given document provides the overview of the proposed recommendation system and a tentative solution by which the optimum services are selected.*

Keywords- collaborative filtering, Recommendation System, optimum services

I. INTRODUCTION

What is big data- Big data is a term for data sets that are so large or complex .big data is having characterized which addresses volume, velocity, and variety, is frequently documented in scientific literature. In recent years, the amount of data in our world has been increasing explosively, and analyzing large data sets—so-called "Big Data"— becomes a key basis of competition underpinning new waves of productivity growth, innovation, and consumer surplus [1]. Then, what is "Big Data"?,Big Data refers to datasets whose size is beyond the ability of current technology, method and theory to capture, man-age, and process the data within a tolerable elapsed time.Today, Big Data management stands out as a challenge for IT companies. The solution to such a challenge is shifting increasingly from providing hardware to provisioning more manageable software solutions [2]. Big Data also brings new opportunities and critical challenges to industry and academia [3] [4].

Similar to most big data applications, the big data tendency also poses heavy impacts on service recommender systems. With the growing number of alternative services, effectively recommending services that users preferred has become an important research issue. Service recommender systems have been shown as valuable tools to help users deal with services overload and provide appropriate recommendations to them. Examples of such practical applications include CDs, books, web pages and various other products now use recommender systems [5], [6], [7]. Over the last decade, there has been much research done both in industry and academia on developing new approaches for service recommender systems [8], [9]

Volume is the most obvious of the three, referring to the size of the data. The massive volumes of data that we are currently dealing with has required scientists to rethink storage and processing paradigms in order to develop the tools needed to properly analyze it[2], [7].

Velocity addresses the speed at which data can be received as well as analyzed [1], [7].

Variety addresses type and nature of the data. This helps people who analyze it to effectively use the resulting insight.

Need of big data: big data has increased the demand of Information management .Developed economies increasingly use data-intensive technologies .According to one estimate, one third of the globally stored information is in the form of alphanumeric text and still image data which is the format most useful for most big data applications. This also shows the potential of yet unused data (i.e. in the form of video and audio content).Data analysis often requires multiple parts of government (central and local) to work in collaboration and create new and innovative processes to deliver the desired outcome. Big data provides an infrastructure for transparency in manufacturing industry, which is the ability to unravel uncertainties such as inconsistent component performance and availability. Predictive manufacturing as an applicable approach toward near-zero downtime and transparency requires vast amount of data and advanced prediction tools for a systematic process of data into useful information [2]. A conceptual framework of predictive manufacturing begins with data acquisition where different type of sensory data is available to acquire such as acoustics, vibration, pressure, current, and voltage and controller data. Vast amount of sensory data in addition to historical data construct the big data in manufacturing.

Survey on Recommendation System:

There is an extensive class of Web applications that involve predicting user responses to options. Such System is called a recommendation system. The system provides a ubiquitous framework with web service technologies and can be applied easily to web and mobile applications.

Social recommendation systems (SRSs) aim to filter the most attractive and relevant content to social media users. Content-based and collaborative filtering are fundamental approaches and are widely adopted for building SRSs[5].Recommendation system that provides users with recommendations of people to invite into their explicit enterprise social network. The recommendations are based on aggregated information collected from various sources across the organization. However, to bring the problem into focus, two good examples of recommendation systems are [1]:

- Offering news articles to on-line newspaper readers, based on a prediction of reader interests.
- Offering customers of on-line retailer suggestions and advices about what they like to buy based on their past history of purchases and/or product searches.

The World Wide Web Consortium (W3C) defines a "Web service" as a software system designed support interoperable machine-to-machine interaction over a network, typically conveyed using hypertext transfer protocol (HTTP) . Representational state transfer (REST) is one such web service architecture for distributed systems .REST-style architectures conventionally consist of clients and servers. Clients initiate requests to servers; servers process requests and return appropriate responses. RESTful concepts to reduce both data transfer and redundant web processing[4].

The recommendation system mainly uses four filters for information retrieval
1. Demographic recommendation system,
2. Content based recommendation system
3. Collaborative recommendation system
4. Hybrids recommendation system

1. Demographic recommendation system -Demographic information can be used to identify the types of users that like a certain object. For example, In this work, we consider an alternative approach to obtaining demographic information in which in which we minimize the effort required to obtain information about user by leveraging the work the user has already expanded in creating a home page on the World Wide Web. Therefore, instead of using approaches to learning from a structured database, we also use text classification to classify users. The positive examples are the HTML home pages of users that like a particular restaurant and the negative examples are the HTML home pages of users that do not like that restaurant. The Winnow algorithm can be used to learn the characteristics of home pages associated with users that like a particular restaurant. We ran an experiment in the exact same manner as the previous experiments. There were six users that did not have or report home pages and we used the text File not found for these home pages. On average, 57.7% of the restaurants in the top three restaurants recommended using a demographic profile were actually liked by the user. **Shaozhong Zhang *et al***[6] proposed Common terms in the profiles included words that referred to the ethnicity of the user or the users home town. While this precision is not as high as other methods, there is an increase over randomly guessing and that information may be combined with other information exploited to increase the precision opredictions Obtaining demographic information can be difficult. LifeStyle Finder enters into dialog with the user to help categorize the user[4], [5].

2. Content-based systems(figure 1) examine properties of the items recommended. For instance, if a Netflix user has watched many cowboy movies, then recommend a movie classified in the database as having the “cowboy” genre**.M. Bjelica *et al***[8] proposed that Recommended systems are a special type of information filtering systems. Information filtering deals with the delivery of items selected from a large collection that the user is likely to find interesting or useful and can be seen as a classification task. Based on training data a user model is induced that enables the filtering system to classify unseen items into a positive classc(relevant to the user) or a negative class c(irrelevant to the user). The training set consists of the items that the user found interesting. These items form training instances that all have an attribute. This attribute specifies the class of the item based on either the rating of the user or on implicit evidence. Formally, an item is described as a vector. The components can have binary, nominal or numerical attributes and are derived from either the content of the items or from information about the users' preferences. The task of the learning method is to select a function based on a training set of minput vectors that can classify any item in the collection. The function ()Xhwill either be able to classify an unseen item as positive or negative at once by returning a binary value or return a numerical value. In that case a threshold can be used to determine if the item is relevant or irrelevant to the use- Content Page- Navigation page- Hybrid Page- Hypertext Link To provide content-based predictions we treat the prediction task as a text-categorization problem. We view movie content information

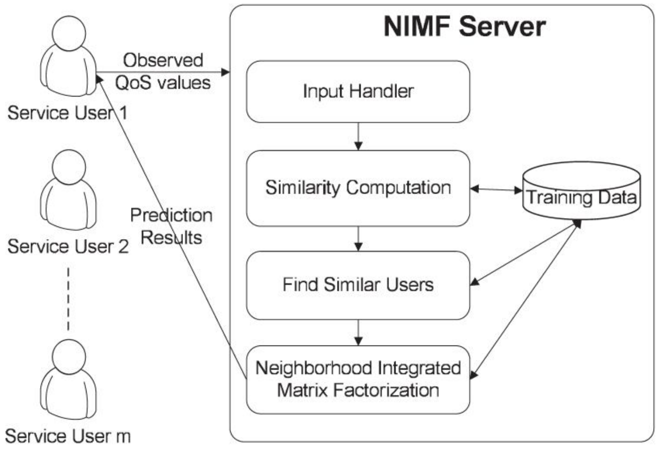

Fig. 1 Content-based systems

text documents, and user ratings 0-5 as one of six class labels. We implemented a bag-of-words naive Bayesian text classifier (Mitchell 1997) extended to handle a vector of bags of words; where each bag-of-words corresponds to a movie-feature (e.g. title, cast, etc.). We use the classifier to learn a user profile from a set of rated movies labeled documents. The learned profile is then used to predict the label (rating) of unrated movies. A similar approach recommending has been used effectively in the book-recommending system. The main work of the content based filter is item oriented similarity, it provides widely desired in the past.

The collaborative filter (figure 2)-The collaborative recommended system only focuses on the user preference of active user. Knowledge based recommendation system means asking user about their own preference of the item and find out what active user required. 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. Collaborative methods look for similarities between users to make predictions. Typically, the pattern of ratings of individual users is used to determine similarity. Such a correlation is most meaningful when there are many objects rated in common between users. For example, in our experimental situation, half of each users ratings are used in the training data. Because the training data is selected from a uniform distribution, on average one quarter of the restaurants will be rated by both users. Since there are 58 restaurants in our sample, approximately 15 are used as the basis of this correlation. In some real situations, we'd expect there to be a smaller number of ratings in common. CF systems work by collecting user feedback in the form of ratings for items in a given domain and exploit similarities and differences among profiles of several users in determining how to recommend an item. The purpose of these dynamic hyper links is to make it easier for a user to find interesting items and thus improving the interaction between the system and the user. .When users browse through a web site they are usually looking for items they find interesting. Interest items can consist of a number of things. For example, textual information can be considered as interest items or an index on a certain topic could be the item a user is looking for. Another example, applicable for a web vendor, is to consider purchased products as interest items. Whatever the items consist of, a website can be seen as a collection of these interest items.

Applications of Recommendation Systems

There are various application of recommendation systems are available, several important applications of recommendation systems are given as:

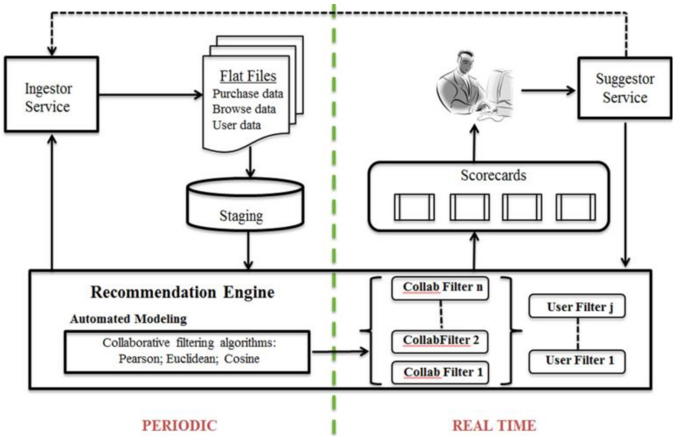

Fig. 2 Collaborative filter

Product Recommendations: The most important use of recommendation systems is at on-line retailers. Noted how Amazon or similar on-line vendors strive to present each returning user with some suggestions of products that they might like to buy. These suggestions are not random, but are based on the purchasing decisions made
by similar customers or on other techniques.

Movie Recommendations(figure3): Netflix offers its customers recommendations of movies they might like. These recommendations are based on ratings provided by users, much like the ratings suggested. The importance of predicting ratings accurately is so high, that Netflix offered a prize of one million dollars for the first algorithm that could beat its own recommendation system by 10%. The prize was finally won in 2009, by a team of researchers called “Bellkor’s Pragmatic Chaos,” after over three years of competition.

News Articles: News services have attempted to

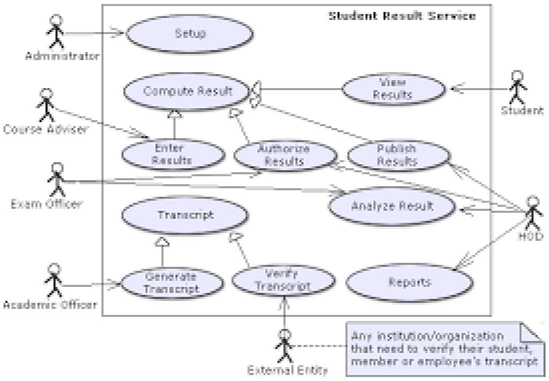

Fig. 3 Recommendation Systems

identify articles of interest to readers, based on the articles that they have read in the past. The similarity might be based on the similarity of important words in the documents, or on the articles that are read by people with similar reading tastes. The same principles apply to recommending blogs from among the millions of blogs available, videos on YouTube, or other sites where content is provided regularly.

Job Recommendation: job recommendation system that aims to assist academia in promoting graduates to appropriate employers. With usability in mind, the system provides a ubiquitous framework with web service technologies and can be applied easily to web and mobile applications. A job recommendation prototype was implemented on the Android and i Phone mobile platforms. Through these systems, users could efficiently seek out and discover appropriate graduates and then recommend them to employers Social recommendation systems (SRSs) aim to filter the most attractive and relevant content to social media users[4].

Big data hadoop(figure 4):-Big data is defined as large amount of data which requires new technologies and architectures so that it becomes possible to extract value from it by capturing and analysis process. Due to such large size of data it becomes very difficult to perform effective analysis using the existing traditional techniques. Big data due to its various properties like volume, velocity, variety, variability, value and complexity put forward many challenges. Since Big data is a recent upcoming technology in the market which can bring huge benefits to the business organizations, it becomes necessary that various challenges and issues associated in bringing and adapting to this technology are brought into light.

II. BACKGROUND

Big DataHadoop/Map reducing:-Hadoop is Google's solution for processing big data and was developed as large Internet search engine providers were the first to truly face the "big data tsunami", indexing billions of WebPages in a quick and meaningful way. Map Reduce is a software framework, written in Java, designed to run over a cluster of machines in a distributed way. The data itself is split into smaller pieces and are distributed over thousands of computers, known as the Google File System (GFS) and a parallelised programming API called Map Reduce is used to distribute the computations to where the data is located (Map) and to aggregate the results at the end (Reduce). Hadoop, an open source implementation of Google's solution, comprised of Map Reduce and the Hadoop Distributed File System (HDFS), is used by leading technology companies such as Face book, Amazon, Twitter and is based on a strategy of co-locating data and processing to significantly accelerate performance. In May 2009, Hadoop broke a world record, sorting a PB of data in 16.25 h and a TB of data in 62 s. Hadoop clusters can be run on private infrastructure, however public offerings such as the Amazon Elastic Map Reduce service (EMR) are proving popular, with EMR enabling users to easily and cost-effectively process large data sets and apply additional analytical techniques such as data mining, machine learning and statistical analysis. It should be noted however that programming Hadoop is not a trivial task; requiring significant expertise in Java to develop parallelised programs. As such, Hadoop has largely only been embraced in the technology sector.

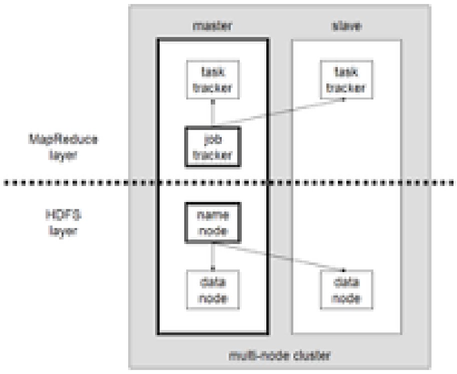

Fig. 4 Big data hadoop

Still evolving at an extremely rapid pace, the application of this technology is now being considered in order to make big data discoveries outside of the technology sector. Given that this platform has now evolved into a widely supported and powerful framework for parallelisation and distribution; two paradigms that are particularly applicable to large genomics and medical data sets, Hadoop has enormous potential for making medical discoveries, if and when applied to the life sciences. Furthermore, given that public clouds such as AWS are now offering DaaS by providing a repository of public data sets, including GenBank, EnsemblE, 1000 Genomes, Model Organism Encyclopaedia of DNA Elements, Unigene, Influenza Virus, this potential is becoming an imminent reality. Hadoop Distributed File System, Hadoop and HBase for genomics, with any improvements contributed back to the open source community. Importantly, Cloudera have teamed up with the Institute for Genomics and Multiscale Biology at the Mount Sinai School of Medicine, in a pioneering effort to aid researchers in applying big data technologies in the field of genomics and multi scale biology to "diagnose, understand and treat disease". Areas of research include analysis of human and bacterial genomes; study of the metabolic pathways of normal and disease states in the organism; structure and function of molecules used in treatment of disease, and more . What is particularly noteworthy about this announcement is that both these organisations are committed to cutting edge research in their fields, and together form a formidable collaboration.MapReduce is a framework for processing parallelizable problems across huge datasets using a large number of computers (nodes), collectively referred to as a cluster (if all nodes are on the same local network and use similar hardware) or a grid (if the nodes are shared across geographically and administratively distributed systems, and use more heterogenous hardware). Processing can occur on data stored either in a filesystem (unstructured) or in a database (structured). MapReduce can take advantage of the locality of data, processing it near the place it is stored in order to reduce the distance over which it must be transmitted.

"Map" step: Each worker node applies the "map()" function to the local data, and writes the output to a temporary storage. A master node ensures that only one copy of redundant input data is processed.

"Shuffle" step: Worker nodes redistribute data based on the output keys (produced by the "map()" function), such that all data belonging to one key is located on the same worker node.

"Reduce" step: Worker nodes now process each group of output data, per key, in parallel.

MapReduce allows for distributed processing of the map and reduction operations. Provided thateach mapping operation is independent of the others, all maps can be performed in parallel – though in practice this is limited by the number of independent data sources and/or the number of CPUs near each source. Similarly, a set of 'reducers' can perform the reduction phase, provided that all outputs of the map operation that share the same key are presented to the same reducer at the same time, or that the reduction function is associative. While this process can often appear inefficient compared to algorithms that are more sequential (because multiple rather than one instance of the reduction process must be run), MapReduce can be applied to significantly larger datasets than "commodity" servers can handle – a large server farm can use MapReduce to sort a petabyte of data in only a few hours.[12] The parallelism also offers some possibility of recovering from partial failure of servers or storage during the operation: if one mapper or reducer fails, the work can be rescheduled – assuming the input data is still available

III. LITERATURE SURVEY

The presented work is to explore the techniques of web usage mining and recommendation system design. Thus this section includes the different techniques are algorithms that are recently introduced for enhancing the application recommendation.Web service is a kind of self-describing programmable application, which help to achieve inter-operability in network environments . It is implemented in a standard language or a with the help of specific protocol web services can be used .The study on Quality-of-Service (QoS) which is a most fundamental element in using web services describing the non-functional characteristics of web services, called boosting. In QoS involves price, response time, throughput and failure rate which perfectly deploy and captures interaction behavior between users and services. With tremendous adoption in the vigorous Internet, sufficient QoS recourse benefits multi- disciplines in web service domain. For example during the process of selection of services, consumer must enjoy massive QoS data, among which it will select desirable subsets through comprehensive comparatative study.

Web Usage mining involves three main steps: Pre-processing, Knowledge Discovery and Pattern Analysis. The information gained from the study and analysis is processed and can be used by the website administrators for efficient dispensation and rationalization of their websites and thus the specific needs of specific communities of users can be fulfilled and profit can be increased. Also, Web Usage Mining help to discover the hidden patterns underlying the Web Log Data. These patterns representing user browsing and navigation behaviors in accessing internet which can be employed in detecting deviations in user browsing behavior in web based banking and other applications where data privacy and security is major concern. ***G. Shivaprasad et al [2]*** Propose pre-process, discovers and analyses the Web Log Data of Dr. T.M.A.PAI polytechnic website. A neuro-fuzzy based hybrid model is employed for Knowledge Discovery from web logs.
Now a day's browsing web has become an important part of our life style. The flowing of data inside the web is very easy and speedy which results the web has become the Centre of data transactions. The various users have unpredictable experience in web. So we can say that the web is a personal experience for each user. Hence need to personalize the web

according to the view of each user making the web a personal experience. ***Sneha Prakash et al [3]*** do this web personalization with the help of a technique called web usage mining. Web usage mining means analyzing the data generated by web surfer's sessions or behaviors. The behavior of a surfer is the click stream data generated by the surfer.
The growth of World Wide Web is incredible as it can be seen in present days. Web usage mining plays an important role in the personalization of Web services, adaptation of Web sites, and the improvement of Web server performance. It applies data mining techniques to discover Web access patterns from Web log data. In order to discover access patterns, Web log data should be reconstructed into sessions. ***Neha Sharma et al [4]*** provide a novel approach for session identification.

The major problem of many online web sites is the presentating and showing many attractives choices to the client at a time; this usually results to tedious and time consuming task in finding the right product or information on the site. In this work, ***D.A. Adeniyi et al [5]*** present a study of Automatic web usage data mining and recommendation system based on current user behavior through his/her click stream data on the newly developed Really Simple Syndication (RSS) reader website, in order to provide relevant information to the user without explicitly asking for it and provide efficient results. The K-Nearest-Neighbor (KNN) classification method has been trained to be used on-line and in Real-Time to identify clients/visitors click stream data, matching it to a particular user group and recommend a tailored browsing option that meet the need of the specific user at a particular time. **Marco A.S. Netto *et al [9]*** proposed that achieve this, web users RSS address file was extracted, cleansed, formatted and grouped into meaningful session and data mart was developed. Our result shows that the K-Nearest Neighbor classifier is transparent, consistent, straightforward, simple to understand, high tendency to possess desirable qualities and easy to implement than most other machine learning techniques specifically when there is little or no prior knowledge about data distribution.

Quality-of-Service (QoS), the most fundamental aspect of web service which attracted numerous customer attention in industry and academia. The analysis on QoS data keeps advancing the state in Service-Oriented Computing (SOC) area. To collect a large amount of resource in practice, QoS prediction applications are designed and built. Nevertheless, how to generate accurate results in high productivity is still a main challenge to existing framework[1],[2].

Users' similarity mining in mobile e-commerce systems is an important field with wide applications, such as personalized recommendation and accurate advertising. Moving trajectories of e-commerce users contain much useful information, providing a very good opportunity for understanding the users' interesting and discovering the similarity between mobile-device-holders. ***Haidong Zhong et al [6]*** explores the problems in the existing mobile ecommerce recommendation methods, and propose a mobile user' moving trajectories mining based user similarity discovering approach for mobile e-commerce system. We formally define the moving trajectory and view the areas, where users stay within for a certain time, as interested regions, which reflect the preferences of mobile-device-holders. Based on the number of overlapped interested areas, a user similarity measure method is proposed. Experimental evaluation, conducted based on the publicly available datasets commendably demonstrate the effectiveness of our approach.Social recommendation systems (SRSs) aim to filter the most attractive and relevant content to social media users. Content-based and collaborative filtering are fundamental approaches and are widely adopted for building SRSs.REST-style architectures conventionally consist of clients and servers. Clients initiate requests to servers; servers process requests and return appropriate responses. Requests and responses are built around the transfer of representations of resources. A resource can essentially be any addressable, coherent, or meaningful concept. A representation of a resource is typically a document that captures the current or intended state of a resource 7 . JavaScript Object Notation (JSON) 8 is a text-based open standard designed for human-readable data interchange

IV. PROBLEM DOMAIN

The proposed work is motivated from the research article given in [7]. In this given process the huge amount of data (big data) and the big data services are used for cluster based recommendation engine design. Basically the recommendation system needs two different entities for correlating them. Therefore the clustering process only provides the seeds for the user recommendations. In addition of that if the clustering technique search a solution in N dimensional space and also theproblem space has the N dimensions. Thus the space complexity of the linear search becomes $O(n^2)$. Therefore the available solution is computationally cost effective for any kind of processing.

V. PROPOSED METHODOLOGY

In order to enhance the solution for computational ability and the for the optimization of recommendation system a quantum genetics based solution is proposed in this work. The proposed methodology of the proposed technique is demonstrated using the figure 1.In the given system the two different kinds of inputs are accepted namely the list of requirements and second the list of services offered by the service providers. In this phase the given problem space is extended by the quantization process and a number of solutions are generated for the different number of problems. Finally the problem and solution strings are encoded in the binary string. These strings are working as the population of the genetic algorithm. The genetic algorithm applies the operators and genetic processes optimize the solutions for minimizing the execution cost of the service offered for the end user and maximize the user requirements for more suitable service selection. The given approach is a formal process of the proposed solution.

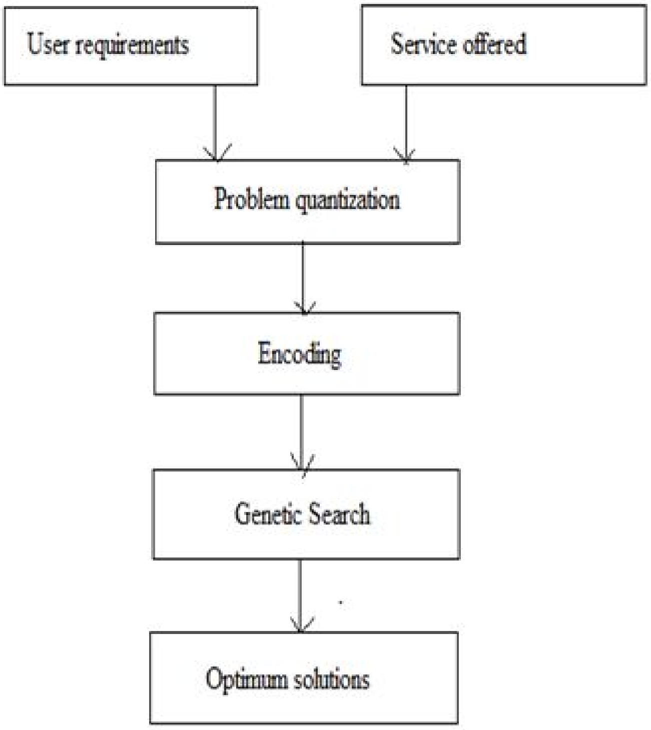

Fig. 5 proposed system

VI. APPLICATION DOMAIN

The proposed solution can be used in various areas of applications some essential application areas are given as:
1. Optimization problems
2. Scheduling problems
3. Production units and resource management

VII. CONCLUSION

In this paper we concentrated on implementation of the proposed recommendation engine which helps the choose best services. A new recommendation engine for web service selection consists of novel optimization technique that maximize the solution space and minimize the problem space by efficient manner of recommendation. Finally, the experimental results demonstrate that user requirments and provides services to improves the accuracy and scalability of service recommender systems over existing approaches.

REFERENCES

[1] "Recommendation Systems", http://infolab.stanford.edu/~ullman/mmds/ch9.pdf

[2] KASR: A Keyword-Aware Service Recommendation Method on MapReduce for Big Data Applications Shunmei Meng, Wanchun Dou, Xuyun Zhang, Jinjun Chen, Senior Member, IEEE

[3] G. Shivaprasad, N. V. Subba Reddy, U. Dinesh Acharya and Prakash K. Aithal, "Neuro-Fuzzy Based Hybrid Model for Web Usage Mining", Eleventh International Multi-Conference on Information Processing-2015 (IMCIP-2015) Published by Elsevier B.V

[4] Sneha Prakash, "Web Personalization using web usage mining: applications, Pros and Cons, Future", International Journal of Computing Science and Information Technology, 2015, Vol.3,Iss.3, 18-26

[5] Ubiquitous job recommendation system for graduates in taiwan,International Journal of Electronic Commerce Studies Vol.6, No.1, pp.127-136, 2015 doi: 10.7903/ijecs.139

[6] Neha Sharma & Pawan Makhija, "Web usage Mining: A Novel Approach for Web user Session Construction", Global Journal of Computer Science and Technology: E Network, Web & Security Volume 15 Issue 3 Version 1.0 Year 2015

[7] D.A. Adeniyi, Z. Wei, Y. Yangquan, "Automated web usage data mining and recommendation system using K-Nearest Neighbor (KNN)classification method", Saudi Computer Society, King Saud University, Applied Computing and Informatics, 2015 Production and hosting by Elsevier B.V

[8] Haidong Zhong, Shaozhong Zhang, Yanling Wang, ShifengWeng and YonggangShu, "Mining Users' Similarity from Moving Trajectories for Mobile Ecommerce Recommendation", International Journal of Hybrid Information TechnologyVol.7, No.4 (2014), pp.309-320

[9] User Based Personalized Search with Big Data, International Journal of Science and Research (IJSR)ISSN (Online): 2319-7064 Index Copernicus Value (2013): 6.14

[10] CloudRec: a framework for personalized service Recommendation in the Cloud Received: 8 February 2013 / Revised: 30 September 2013 / Accepted: 04 December 2013 / Published online: 9 January 2014 © Springer-Verlag London 2014