A Novel Information Retrieval System for Effective Acquisition of Data using Cloud Computing

R. Malathi Ravindran[1], Dr.Antony Selvadoss Thanamani[2]
[1, 2] Department of Computer Science
[1, 2] NGM College, Pollachi

Abstract- *With the development of information technologies, data volumes processed by many applications will routinely cross the peta scale threshold which would in turn increase the computational requirements. The existing information retrieval like Boolean, meta and probabilistic have the inverse correlation in precision and recall measures. It is also difficult to retrieve the data when the size of the data increases [2]. So in this work, in order to retrieve the needed information from the large volume of the data, a cloud based information retrieval system is proposed with the inclusion of vector space model and semi supervised clustering. In this work, Dow Jones Index dataset is experimented using math work. The result proves that the Information Retrieval System in Cloud performs better than the native mechanisms.*

Keywords- Information Retrieval System, Dow Jones Index dataset, Vector Space Model, Cloud Computing, Semi Supervised Clustering.

I. INTRODUCTION

Information Retrieval is a thrust area of research. Several methodologies are adopted and adapted to retrieve information. Information retrieval is a discipline involved with the organization, storage, retrieval, and display of bibliographic information. Information retrieval (IR) is finding material (usually documents) of an unstructured nature (usually text) that satisfies an information need from within large collections (usually stored on computers)[1]. There are lot of search engines available, including web search engines, selection-based search engines, meta search engines, desktop search tools, and web portals and vertical market websites that have a search facility for online databases. Some of the search engines are Privacy search engines, Open source search engines, Semantic browsing engines, Social search engines, Visual search engines, Search appliances and Desktop search engines. The limitations of the existing search engine models are poor precision, poor recall, varied document quality and varied indexing depth [8].

Since the information is moving towards bigdata, the existing retrieval system is not suitable to access the data effectively. So in this research work, a solution is proposed for retrieving the objects and processes in the context of semi supervised clustering using advanced vector space mode. The proposed Information Retrieval System in Cloud is validated using math works.

II. EXISTING INFORMATION RETRIEVAL METHOD

Due to widely use computer, Information Retrieval from various cloud service providers via internet is becoming more important because of the fact that the best, particularly the quickest conveniently available source of knowledge on web source. Therefore, web search engines are rising as very useful and reliable tools in knowledge finding and research activities.

A. BOOLEAN

The interest for information retrieval has existed long before the Internet. The boolean retrieval is the most simple of these retrieval methods and relies on the use of Boolean operators. The terms in a query are linked together with AND, OR and NOT. This method is often used in search engines on the Internet because it is fast and can therefore be used online. This method has also its problems. The user has to have some knowledge to the search topic for the search to be efficient, e.g., a wrong word in a query could rank a relevant document non relevant. The retrieved documents are all equally ranked with respect to relevance and the number of retrieved documents can only be changed by reformulating the query.

The Boolean retrieval has been extended and refined to solve these problems. Expanded term weighting operations make ranking of documents possible, where the terms in the document could be weighted according to their frequency in the document. Boolean information retrieval has been combined with content-based navigation using concept lattices, where shared terms from previously attained documents are used to refine and expand the query. The Boolean operators have been replaced with fuzzy operators. Weighted query expansion using a thesaurus. A model based on fuzzy set theory allows the interpretation of a user query with a linguistic descriptor for each term.

B. META

A meta search engine (or aggregator) is a search tool that uses another search engine's data to produce their own

results from the Internet. Meta search engines take input from a user and simultaneously send out queries to third party search engines for results. Sufficient data is gathered, formatted by their ranks and presented to the users.

Information stored on the World Wide Web is constantly expanding, making it increasingly impossible for a single search engine to index the entire web for resources. A meta search engine is a solution to overcome this limitation. By combining multiple results from different search engines, a meta search engine is able to enhance the user's experience for retrieving information, as less effort is required in order to access more materials. A meta search engine is efficient, as it is capable of generating a large volume of data, however, scores of websites stored on search engines are all different: this can draw in irrelevant documents. Other problems such as spamming also significantly reduce the accuracy of the search. The process of fusion aims to tackle this issue and improve the engineering of a meta search engine. There are many types of meta search engines available to allow users to access specialised information in a particular field. These include Savvysearch engine and Metaseek engine. The architecture of meta search engine is as follows;

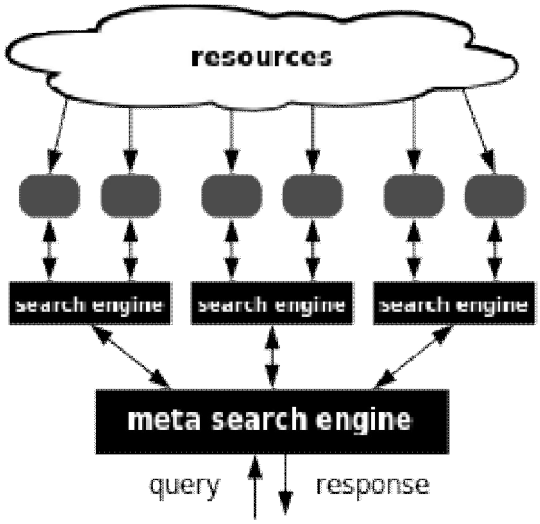

Figure1: Architecture of Meta search Engine

A meta search engine accepts a single search request from the user. This search request is then passed on to another search engine's database. A meta search engine does not create a database of web pages but generates a virtual database to integrate data from multiple sources. Since every search engine is unique and has different algorithms for generating ranked data, duplicates will therefore also be generated. To remove duplicates a meta search engine processes this data and applies its own algorithm. A revised list is produced as an output for the user. When a meta search engine contacts other search engines, these search engines will respond in three ways:

They will both cooperate and provide complete access to interface for the meta search engine, including private access to the index database, and will inform the meta search engine of any changes made upon the index database;

- Search engines can behave in a non-cooperative manner whereby they will not deny or provide any access to interfaces;
- The search engine can be completely hostile and refuse the meta search engine total access to their database and in serious circumstances, by seeking legal methods.

C. PROBABILISTIC

Another classic retrieval method is the probabilistic retrieval, where the probability that a specific document will be judged relevant to a specific query, is based on the assumption that the terms are distributed differently in relevant and non-relevant documents. The probability formula is usually derived from Bayes' theorem [9]. Expansion of the Probabilistic retrieval model is to incorporate relationships of the document descriptors.

III. LITERATURE SURVEY

In [1] ChengXiangZhai, discussed the Statistical language models have recently been successfully applied to many information retrieval problems. A great deal of recent work has shown that statistical language models not only lead to superior empirical performance, but also facilitate parameter tuning and open up possibilities for modeling non-traditional retrieval problems. In general, statistical language models provide a principled way of modeling various kinds of retrieval problems. The purpose of this survey is to systematically and critically review the existing work in applying statistical language models to information retrieval, summarize their contributions, and point out outstanding challenges. In [4] Jacek Gwizdka and Mark Chignell denote that the Information retrieval on the Web is very different from retrieval in traditional indexed databases. This difference arises from: the high degree of dynamism of the Web; its hyper-linked character; the absence of a controlled indexing vocabulary; the heterogeneity of document types and authoring styles; the easy access that different types of users may have to it. Thus, since Web retrieval is substantially different from information retrieval, new or revised evaluative measures are required to assess retrieval performance using Web search engines. In the second part of the paper, application of these measures is illustrated in the evaluation of three search engines. The purpose of this paper is not to give the definite prescription for evaluating information retrieval

from the Web, but rather to present some examples and to initiate a wider discussion of how to enhance measures of Web search performance. In [5] Madhuri.H Parekh describes Cloud computing is the latest technology that delivers computing resources as a service such as infrastructure, storage, application development platforms, software etc. Huge amount of data is stored in the cloud which needs to be retrieved efficiently. In Cloud Computing using of Clustering Process from Heterogeneous Network fetch the data find out the row data. The retrieval of information from cloud takes a lot of time as the data is not stored in an organized way. Data mining is thus important in cloud computing. We can integrate data mining and cloud computing which will provide agility and quick access to the technology. The integration should be so strong that it will be able to deal with increasing production of data and will help in efficient mining of massive amount of data. In this paper, we provide brief description about cloud computing and clustering techniques. Then, it also describes about cloud data mining. This paper proposes a model that applies traditional hierarchical improved agglomerative clustering algorithm and distributed on heterogeneous network. In [8] Shilpy Sharma explores as the web continues to grow exponentially, the idea of crawling the entire web on a regular basis becomes less and less feasible, so the need to include information on specific domain, domain-specific search engines was proposed. This paper describes the use of semi-structured machine learning approach with Active learning for the "Domain Specific Search Engines". A domain-specific search engine is "An information access system that allows access to all the information on the web that is relevant to a particular domain. The proposed work shows that with the help of this approach relevant data can be extracted with the minimum queries fired by the user. It requires small number of labeled data and pool of unlabelled data on which the learning algorithm is applied to extract the required data.

IV. PROPOSED INFORMATION RETRIEVAL MODEL

The field of information retrieval is the study of methods to provide users with that small subset of information relevant to their needs and to do so in a timely fashion[6]. An information retrieval system is a technique that provides solutions to manage information on documents. The system assists users in finding the information they need. An object is an entity, which is represented as information in the database. The queries given by the users are matched against the database information. Objects may be in various types such as images, text documents, videos and audios. The information retrieval process begins as soon as user enters a query into the system. Queries not only extract a single unique content from the data source instead of that it collects multiple answers which is almost relevant to the given query. These documents are called relevant documents. A perfect retrieval system would retrieve only the relevant documents and no irrelevant document. However, perfect retrieval systems do not exist and will not exist because search statements are necessarily incomplete and relevance depends on the subjective opinion of the user. Two users may pose the same query to an information retrieval system and give different relevance judgments on the retrieved documents.

In general documents does not store directly into the information retrieval system, instead of that it stores with metadata. Information retrieval system matches the user given query with database and extracts the result and display based on the rank basis. Highly matched results will be listed at first and lower matched result will be listed at last, remaining results will be listed between these two based on the ranking.

Information retrieval is a major area of research. There is no efficient system available for information retrieval. To overcome the inconsistencies this research work introduced effective solution for information retrieval objects and processes in the context of semi-supervised clustering using advanced vector space model. More importantly this research work proposed an advanced vector space based efficient information retrieval system.

The vector space model [3][7] provides the framework for most information retrieval algorithms used today. However, this most basic vector space model alone is not efficient enough. Many modifications required to speed up the basic model. In this research work advanced vector space model, cloud computing along with semi-supervised clustering has been introduced to develop efficient information retrieval system in Cloud.

A. CLOUD ARCHITECTURE

Cloud computing architecture refers to the components and subcomponents required for cloud computing. These components typically consist of a front end platform (fat client, thin client, mobile device), back end platforms (servers, storage), a cloud based delivery, and a network (Internet, Intranet, Intercloud). Combined, these components make up cloud computing architecture. Cloud computing refers to the use of computers which access Internet locations for computing power, storage and applications, with no need for the individual access points to maintain any of the infrastructure. The following figure shows the architecture of cloud computing.

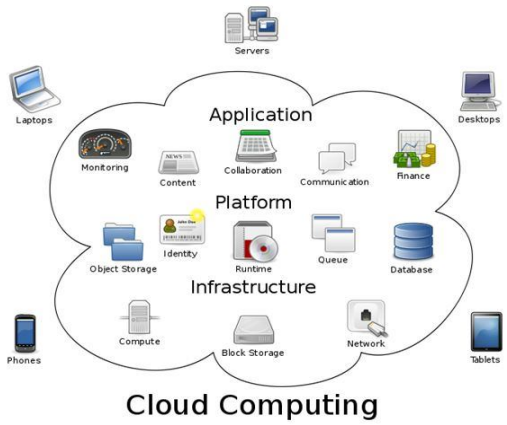

Figure 2: Cloud Computing Architecture

To shed light on the advantages and disadvantages of cloud computing's three main service delivery models - software-as-a-service (SaaS), platform-as-a-service (PaaS) and infrastructure-as-a-service (IaaS) [10].

1. INFRASTRUCTURE-AS-A-SERVICE

The IaaS delivery model represents a self-contained IT environment comprised of infrastructure-centric IT resources that can be accessed and managed via cloud service-based interfaces and tools. This environment can include hardware, network, connectivity, operating systems, and other "raw" IT resources. In contrast to traditional hosting or outsourcing environments, with IaaS, IT resources are typically virtualized and packaged into bundles that simplify up-front runtime scaling and customization of the infrastructure.

The general purpose of an IaaS environment is to provide cloud consumers with a high level of control and responsibility over its configuration and utilization. The IT resources provided by IaaS are generally not pre-configured, placing the administrative responsibility directly upon the cloud consumer. This model is therefore used by cloud consumers that require a high level of control over the cloud-based environment they intend to create.

Sometimes cloud providers will contract IaaS offerings from other cloud providers in order to scale their own cloud environments. The types and brands of the IT resources provided by IaaS products offered by different cloud providers can vary. IT resources available through IaaS environments are generally offered as freshly initialized virtual instances. A central and primary IT resource within a typical IaaS environment is the virtual server. Virtual servers are leased by specifying server hardware requirements, such as processor capacity, memory, and local storage space [10].

2. PLATFORM-AS-A-SERVICE

The PaaS delivery model represents a pre-defined "ready-to-use" environment typically comprised of already deployed and configured IT resources. Specifically, PaaS relies on (and is primarily defined by) the usage of a ready-made environment that establishes a set of pre-packaged products and tools used to support the entire delivery lifecycle of custom applications.

Common reasons a cloud consumer would use and invest in a PaaS environment include:

- The cloud consumer wants to extend on-premise environments into the cloud for scalability and economic purposes.
- The cloud consumer uses the ready-made environment to entirely substitute an on-premise environment.
- The cloud consumer wants to become a cloud provider and deploys its own cloud services to be made available to other external cloud consumers.

By working within a ready-made platform, the cloud consumer is spared the administrative burden of setting up and maintaining the bare infrastructure IT resources provided via the IaaS model. Conversely, the cloud consumer is granted a lower level of control over the underlying IT resources that host and provision the platform. PaaS products are available with different development stacks. For example, Microsoft Azure provides a .NET-based environment, while Google App Engine offers a Java and Python-based environment [10].

3. SOFTWARE-AS-A-SERVICE

A software program positioned as a shared cloud service and made available as a "product" or generic utility represents the typical profile of a SaaS offering. The SaaS delivery model is typically used to make a reusable cloud service widely available (often commercially) to a range of cloud consumers. An entire marketplace exists around SaaS products that can be leased and used for different purposes and via different terms.

A cloud consumer is generally granted very limited administrative control over a SaaS implementation. It is most often provisioned by the cloud provider, but it can be legally owned by whichever entity assumes the cloud service owner role. For example, an organization acting as a cloud consumer while using and working with a PaaS environment can build a cloud service that it decides to deploy in that same environment as a SaaS offering. The same organization then effectively assumes the cloud provider role as the SaaS-based cloud service is made available to other organizations that act

as cloud consumers when using that cloud service [10].

V. RESULTS

In this research work, mathwork is used to simulate the dataset. The results has been graphically demonstrated with mathwork [5]. Dow Jones Index Data Set is used in this work.

A. Dow Jones Index Data Set

On July 3, 1884, Charles Henry Dow began publishing his Dow Jones Average. By the time it was published daily eight months later, the index was composed of 12 stocks, 10 of which were railroads. This index appeared in the Customer's Afternoon Letter up until July 8, 1889 when the first issue of The Wall Street Journal was published. On October 7, 1896, Dow started publishing two "Daily Moving Averages," 12 industrials and 20 railroads (that would later become the transportation index.) This first Dow Jones Industrial Average (DJIA) was published through September 29, 1916.

This first DJIA closed at 71.42 on July 30, 1914 and so did the New York Stock Market for the next four months. Some historians believe the reason for this was worry that markets would plunge because of panic over the onset of the World War. An interesting book by William L. Silber titled When Washington Shut down Wall Street: The Great Financial Crisis of 1914 and the Origins of America's Monetary Supremacy (2007) has a different explanation. He thinks that Secretary of the Treasury, William McAdoo closed the exchange because he wanted to conserve the US gold stock in order to launch the Federal Reserve System later that year with enough gold to keep the US on the gold standard. Whatever the reason, the first day it reopened on December 12, 1914, the index closed at 74.56, thus the War had not had the predicted impact.

On October 4, 1916, the WSJ starts publishing a (new) DJIA of 20 stocks, 8 stocks from the old index and 12 new stocks. It was traced the index back to December 12, 1914 at that time. It is important to know that data for the first DJIA of 12 stocks and the second DJIA of 20 stocks are BOTH available for the 21 months and the first index was about 36% higher than the second and the data here are adjusted to make them a consistent time series. On October 1, 1928 the DJIA of 30 stocks was first quoted in the WSJ. It contained 14 stocks from the second list of 20 and 16 new stocks. Its level was consistent with the second list, so no adjustment was necessary. Dow Jones Index Data Set is available in University of Maryland University College.

Dow Jones Index dataset contains weekly data for the Dow Jones Industrial Index. It has been used in computational investing research. Dow Jones dataset contains 750 instances. This dataset characteristics is an time-series. Integer and Real are the two attribute characteristics. Clustering and Association are the major associated tasks of this dataset. The area of this dataset is ‘Business’.

Dow Jones Index dataset is experienced with Mathworks and illustrated with clear graphs in following subsections.

This graph drawn based on 23 attributes such as quarter: the yearly quarter(Jan-Mar=1 and Apr-Jun=2), stock, date: the last business day of the work, open: the price of the stock at the beginning of the week, high: the highest price of the stock during the week, low: the lowest price of the stock during the week, close: the price of the stock at the end of the week, volume: the number of shares of stock that traded hands in the week, percent change price: the percentage change in price throughout the week, percent change volume over last week: the percentage change in the number of shares of stock that traded hands for this week compared to the previous week, previous weeks volume: the number of shares of stock that traded hands in the previous week, next weeks volume: the opening price of the stock in the following week, next weeks close: the closing price of the stock in the following week, percent change next weeks price: the percentage change in price of the stock in the following week days to next dividend, percent return next dividend: the percentage of return on the next dividend. Indegree, outdegree, nodel centroid, eccentric centroid, centroid, eccentricity are calculated based on remaining attributes in dataset.

B. Experimental Analysis of Dow Jones Index dataset using Native Mechanisms

The graph shown in Figure 1 illustrates the various aspects of metabolic network for Quater: 1 - Stock: CVX - Date: 21/1/2011.

This graph is drawn by using 23 attributes. In this graph attribute PCV_OLWeek stands for percent_change_volume_over_last_wk, NW_open () stands for next_weeks_open (), NW_close () stands for next_weeks_close (), PC_NW_price stands for percent_change_next_weeks_price, DTN_dividend stands for days_to_next_dividend, PRN_dividend stands for percent_return_next_dividend, Clus_Coefficient stands for ClusteringCoefficient, Bet_Centrality stands for Betweenness Centrality and Close_Centrality stands for Closeness Centrality.

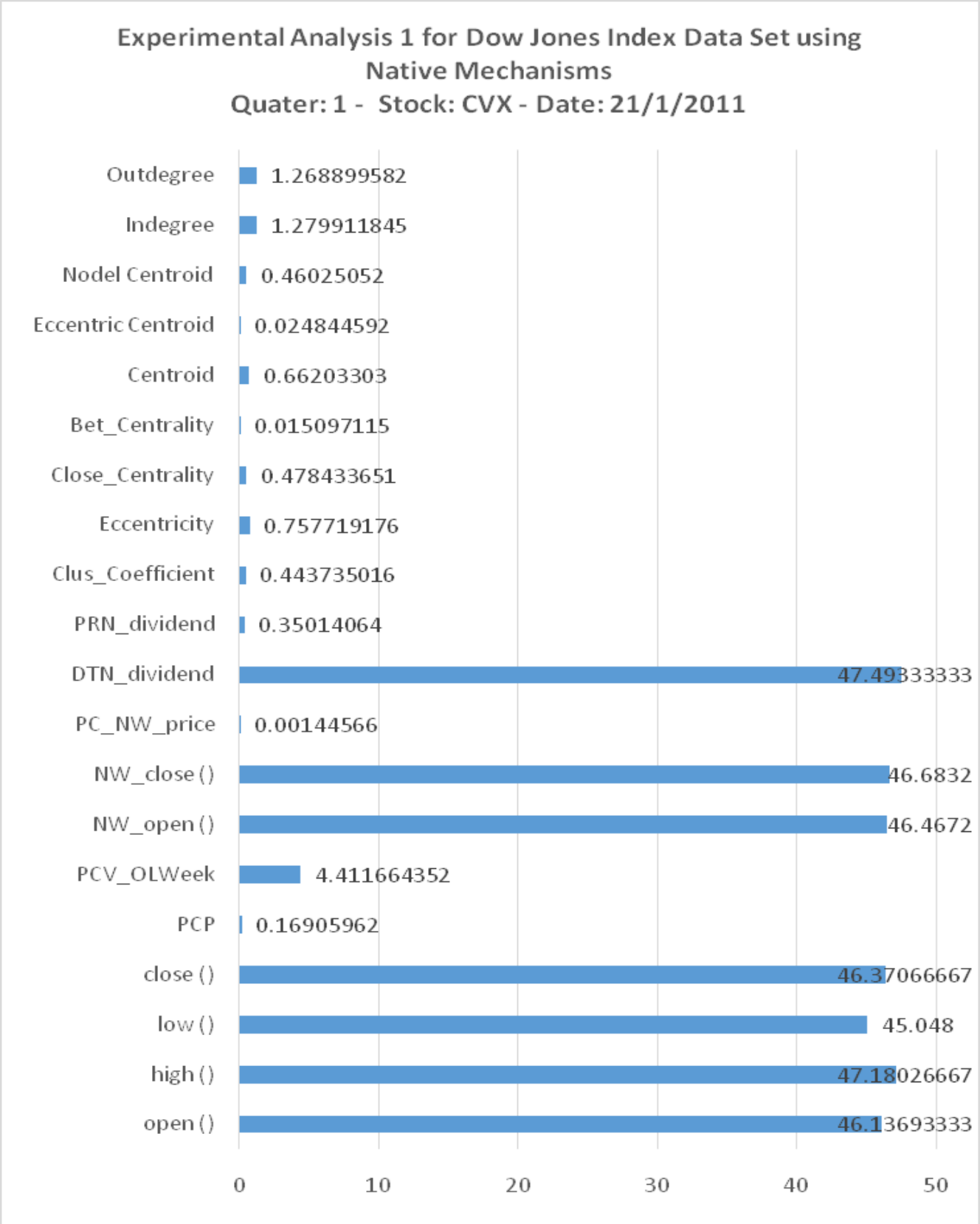

Figure 1: Experimental Analysis 1 for Dow Jones Index Data Set using Native Mechanisms [Quater: 1 - Stock: CVX – Date: 21/1/2011]

C. Average experimental analysis for Dow Jones Indes dataset using native mechanisms

The graph shown in Figure 2 illustrates the various aspects of metabolic network.

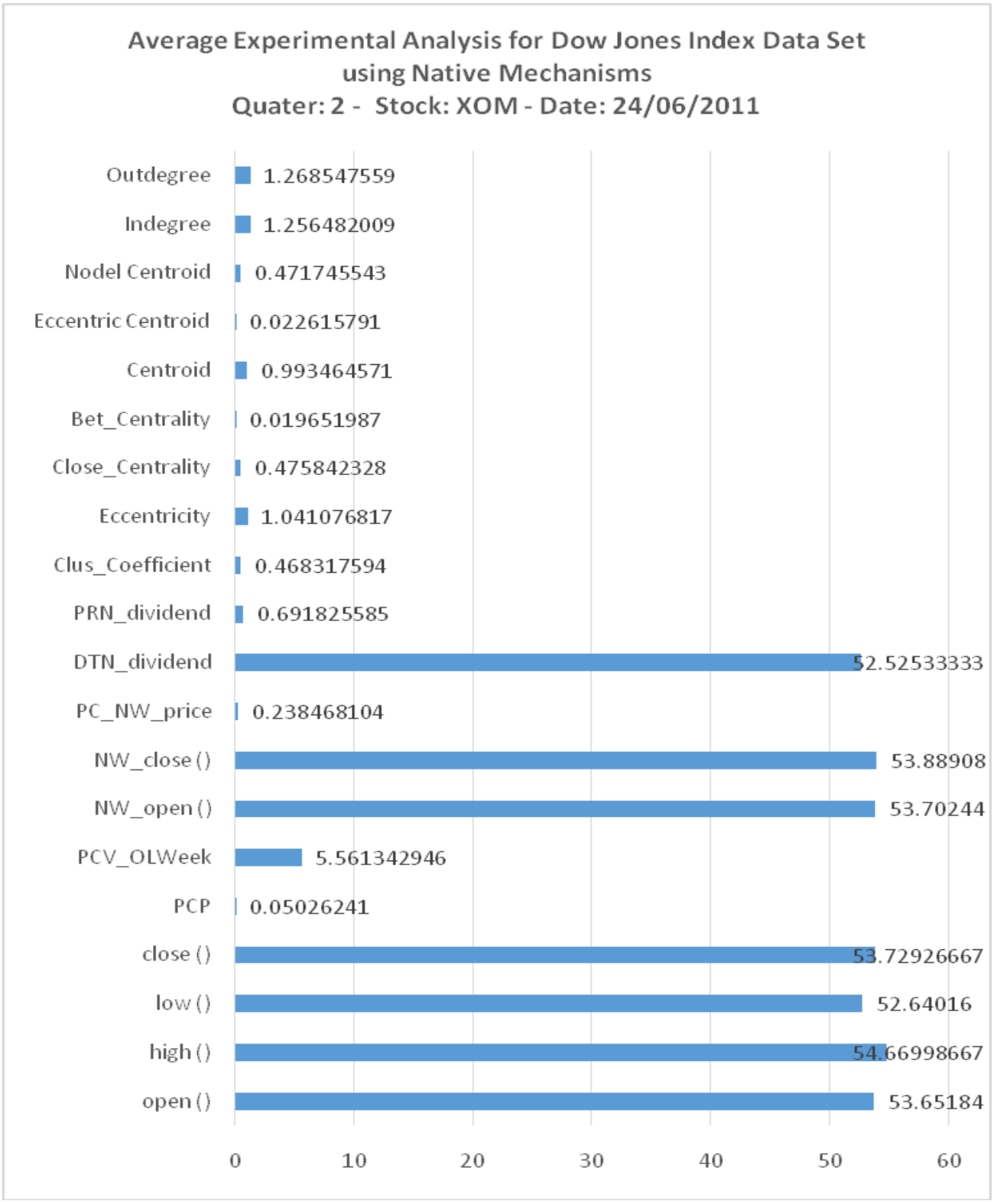

Figure 2: Average Experimental Analysis for Dow Jones Index Data Set using Native Mechanisms [Quater: 2 - Stock: XOM - Date: 24/06/2011]

D. Average Experimental Analysis for Dow Jones Index dataset using Information Retrieval System in Cloud

The graph shown in Figure 3 illustrates the various aspects of metabolic network for average experimental analysis of Quarter: 2 - Stock: XOM - Date: 24/06/2011.

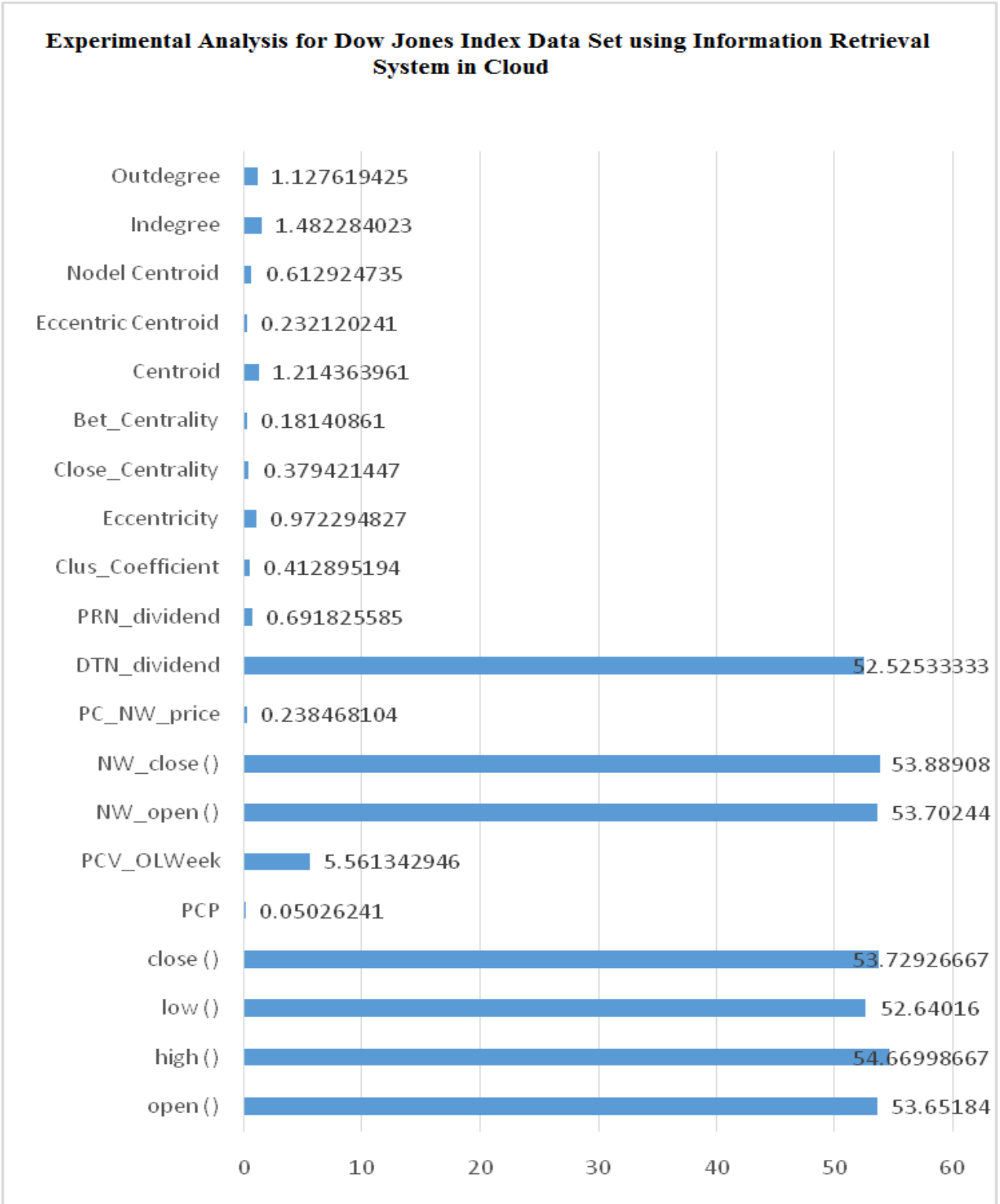

Figure 3: Average Experimental Analysis for Dow Jones Index Data Set using CCBIR System [Quater: 2 - Stock: XOM – Date: 24/06/2011]

This graph drawn by using 23 attributes. In this graph attribute PCV_OLWeek stands for percent_change_volume _over_last_wk, PWV stands for previous_weeks_volume, NW_open () stands for next_weeks_open (), NW_close () stands for next_weeks_close (), PRN_dividend stands for percent_return_next_dividend, Clus_Coefficient stands for ClusteringCoefficient and Bet_Centrality stands for Betweenness Centrality.

VI. CONCLUSION

Information Retrieval System provides solutions to manage information on documents. There is no efficient system available for Information Retrieval. To overcome the inconsistencies, this research work introduce effective solution for Information Retrieval object and processes in the context of semi supervised clustering using advanced Vector Space Model. The proposed information retrieval system works successfully and it is validated using Dow Jones Index data set. It also compares the experimental analysis of Dow Jones Index dataset metabolic relation network using native mechanisms and proposed information retrieval model. The results prove that the proposed model is more efficient than the native mechanisms.

REFERENCES

[1] Chengxiang Zhai, "Statistical Language Models for Information Retrieval A Critical Review", Foundations and Trends in Information Retrieval Vol.2, No.3, pp.137-213, 2008.

[2] David L. Donoho (2000), "High Dimensional Data Analysis: The Curses and Blessings of Dimensionality," American Math. Society Conference: Mathematical Challenges of the 21st Century, Los Angeles, CA, August, 6-11 (2000).

[3] Erk, K., & Pad´o, S. (2008). "A structured Vector Space Model for word meaning in context" in Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing (EMNLP-08) , pp. 897–906, Honolulu, HI.

[4] Jacek Gwizdka and Mark Chignell, "Towrds Information Retrieval Measures for Evaluation of Web Search Engines (1999)", citeseerx.ist.psu.edu.

[5] Madhuri H.Parekh, "Enhancement Clustering of Cloud Datasets using Improved Agglomerative Technique", International Journal of Advanced Networking Applications, pp.128-131.

[6] Riktesh Srivastava, "Implementation of Information Retrieval (IR) Algorithm for Cloud Computing: A Comparative Study between with and without MapReduce Mechanism", Journal of Contemporary Issues in Business Research, Vol.1, No.2, pp.42-56, 2012

[7] G.Salton A.wong et al., "A Vector Space Model for automatic indexing", Communications of the ACM, Vol.18, No.11, pp.613-620. Nov.1975.

[8] Shilpy Sharma, "Information Retrieval in Domain Specific Search Engine with Machine Learning Approaches", World Academy of Science, Engineering and Technology, International Scholarly and Scientific Research & Innovation Vol.2 Issue 6, 2008.

[9] Van Rijbergen.C.J, "Information Retrieval", Butterworths, 1979.

[10] www.whatiscloud.com

BIOGRAPHY

Ms. R.Malathi Ravindran received MCA degree from Madurai Kamarajar University, Madurai, India. She has completed her Master of Philosophy in Periyar University, Salem. Presently she is working as an Assistant Professor of MCA in NGM College, Pollachi. She has 10 years of teaching experience. Her area of interest includes data Mining. Now she is pursuing her Ph.D in Computer Science in the Research Department of Computer Science, NGM College, Pollachi under Bharathiar University, Coimbatore, India.

Dr. Antony Selvadoss Thanamani is presently working as Associate Professor and Head, Research Department of Computer Science, NGM College, Pollachi, Coimbatore, India. He has published more than 100 papers in international/national journals and conferences. He has authored many books on recent trends in Information Technology. His areas of interest include E-Learning, Knowledge Management, Data Mining, Networking, Parallel and Distributed Computing. He has to his credit 25 years of teaching and research experience. He is a senior member of International Association of Computer Science and Information Technology, Singapore and Active member of Computer Science Society of India, Computer Science Teachers Association, New York.