International Journal of Database Theory and Application
Vol.8, No.6 (2015), pp.47-56
http://dx.doi.org/10.14257/ijdta.2015.8.6.05

An Efficient Semantic Ranked Keyword Search of Big Data Using Map Reduce

P. Srinivasa Rao, M.H.M. Krishna Prasad and K. Thammi Reddy

Associate Professor, Professor, Professor
Dept.of CSE, MVGRCE, Dept.of CSE, JNTUK, Dept.of CSE
GITAM University
Vizianagaram, Kakinada, Visakhapatnam
psr.sri@gmail.com, Krishnaprasad.mhm@gmail.com, thammireddy@yahoo.com

Abstract

Information retrieval is fast becoming the prevailing form of information access, surpassing traditional database style searching. Ontologies have become the tool of choice employed in many information retrieval systems and more prominently in semantic information retrieval. In order to overcome the disadvantages in key word based information retrieval systems, which transfer irrelevant information, ontology has been designed. A system with ontology mimics the real world, where every task is laced with certain meaning as this is basic idea behind knowledge processing. Hadoop, which is an open source frame work for storing and processing large datasets, is used for pre-processing the text documents. First, a set of text documents are considered. Pre-processing is performed on a large domain of data using Hadoop MapReduce. This includes the removal of the stop words along with stemming and excluding less frequency words. Despite this pre-processing, owing to the colossal number of index terms still floating in the considered domain data, the problem of high dimensionality is encountered. Therefore the dimensionality of such a group of terms is reduced by identifying it as a concept and those concepts can be viewed as a single dimension in a ontology based information retrieval system. Now ontology is constructed by assigning synonym set to each concept in this structure using tools like word net. Thus constructed ontology can be mapped on to the processed query which gives us the relevant information from the data pool considered.

Keywords: *MapReduce, Bigdata, Hadoop, Datamining, Information Retrieval System*

1. Introduction

Information retrieval is fast becoming the prevailing form of information access, surpassing traditional database style searching. The reason for the need of information retrieval is to process large corpus (the group of documents over which we perform retrieval is called collection or corpus) quickly, to allow more flexible searching operations, to allow ranked retrieval. Presently many search engines and systems are based on keyword based retrieval methodology which has its own limitations regardless of many effective improvements in its retrieving methods. Keyword based retrieval systems face difficulties in conceptualization of user needs. Aiming to solve the limitations of keyword-based models, the idea of semantic search, has been the focus of a wide body of research in the Information Retrieval (IR) and the Semantic Web (SW) communities [10]. At the core of these new technologies, ontologies were envisioned as key elements to overcome the limitations of key word based search [13]. Modern information retrieval systems need the

ISSN: 2005-4270 IJDTA

capability to reason about the knowledge conveyed by text bases. The ontology constructed [11] allows users to query the semantic content of the documents. Ontology is a collection of concepts and their interrelationships [12] which can collectively provide an abstract view of an application domain.

The retrieval of huge information using Ontology can be improved using one powerful tool called Hadoop MapReduce. Representation of document as a document term vector leads to high dimensionality problem as there will be enormous number of index terms in corpus. It is possible to reduce the dimensionality by considering equivalence classes of terms associated with related concepts as single dimension for rough set model. So clustering is used to group the terms. A cluster is described as a set of similar objects or entities collected or grouped together. All entities within a cluster are alike and the entities in different clusters are not alike. Each entity may have multiple attributes, or features and the likeness of entities is measured based on the closeness of their features. Therefore, the crucial point is to define proximity and a method to measure it. There are many clustering techniques and algorithms in use. K-means is the most common and often used algorithm. K-means algorithm takes an input parameter k, and partitions a set of n objects into k clusters according to a similarity measure.

However the results of partition clusters are highly influenced by the initial selection of seed points. It may not generate quality of clusters if it starts with randomly generated seed points.one solution is to make use of multiple set of random seed points as the basis for generation of clusters and selecting the best of clusters base on their quality. This requires a lot of computational resources. We have many paths in which we can carry out all the required computations where Hadoop is one among those paths available which is fast robust easier to understand and relatively efficient to perform the required computations. Hadoop is extremely scalable. Major component of Hadoop HDFS (for storage) is optimized for high through put. Hadoop is an open source software framework that can run large data-intensive, distributed applications and can be installed on commodity Linux clusters. Hadoop comes with its own file system called the Hadoop Distributed File System (HDFS) and a strong infrastructural support for managing and processing huge petabytes of data. Each HDFS cluster consists of one unique server called the Name node that manages the namespace of the system, determines the mapping of blocks to Data nodes, and regulates file access. Each node in the HDFS cluster is a Data node that manages the storage attached to it. The data nodes are responsible for serving read and write requests from the clients and performing block creation, deletion and replication instructions from the Name node.

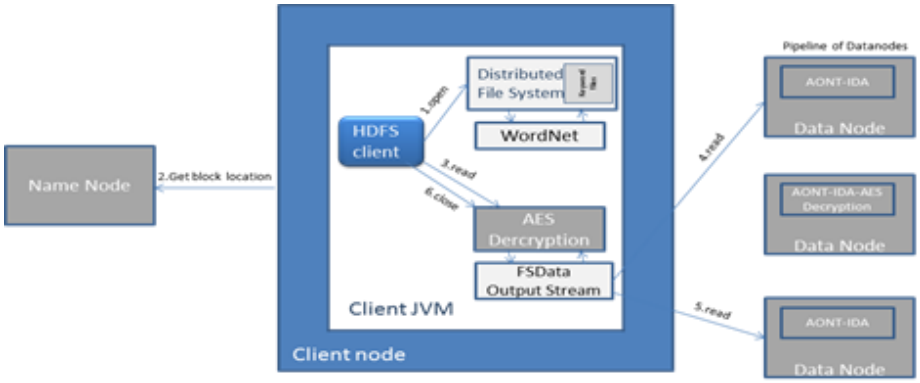


Figure 1. System Model Using Hadoop MapReduce

2. Related Work

Pallawi unmesh bulakh *et al.* [1] concerned with the application of semantic web regarding semantic similarity between the words, documents. Current assessment system where analysis of subjective question papers that can be done manually is discussed. So, there is a need to automate this task by developing software that can automatically extract ontologies from given documents, and perform semantic analysis on it. For this, the answer papers are first represented in ontological form. Here the authors used the lexical database defined by WordNet and the application of semantic similarity algorithms to obtain similarity between documents.

Thomas C. Jepsen [2] describes some definitions of "ontology" as it relates to computer applications and gives an overview of some common ontology-based applications. From a more modern perspective, ontologies came to be of interest to computer scientists in the 1970s as they began to develop the field of artificial intelligence. They realized that if you could create a domain of knowledge and establish formal relationships among the items of knowledge in the domain, you could perform certain types of automated reasoning. Tom Gruber, a computer science scientist introduced the term in his paper in 1993. Thomas also described the properties of ontology, types of ontology and also how it is different from hierarchies.

Marek Obitko, *et al.*. [3] described how to design Ontology using Formal concept analysis. Their ontology design allows for discovering necessity for new concepts and relations, which leads to an ontology which is suitable for knowledge exchange and information retrieval. The main characteristics of their method are: concepts are described by properties. The properties determine the hierarchy of concepts. When the properties of different concepts are same, then the concepts are same as well. However if the context increase, construction and navigation of ontology becomes complex.

Abdelmalek Amine *et al.* [4] proposed three text clustering methods which are examined using 3 similarity measurements in which F-measure evaluation criteria is used. author also explained advantages and disadvantages of each method. The results show that the SOM-based clustering method using the cosine distance provides the better results.

Madalina ZURINI *et al.* [5] Described word sense disambiguation (WSD) for processing documents in order to increase the correctness of the classification. Presented principal distance measures using the graph associated to WordNet .Advantages and disadvantages of measures were explained. Similarity measure based on probabilities of co-occurrences is used for non-existing words in WordNet.

Christos Bouras *et al.* [6] described two Clustering methodologies such as hierarchical clustering and patronal clustering. Author presented an algorithm for enhancing K-Means using WordNet. Different similarity measures were applied on clustering methods. Author concluded that Euclidian distance measure and cosine distance measures on K-Means, produced better results.

K. Thammi Reddy, *et al*. [7] proposed a rough set based information retrieval in which they presented a hybrid clustering approach for the formation of equivalence classes of terms associated with related concepts. They also proposed a new term weight estimate namely term probability–inverse document frequency (TF-IDF) for representing a term as a vector before clustering the terms. Clustering is performed to group together related terms of a concept into equivalence classes, which can be used to reduce the dimensionality of the documents for rough classification.

Jian-Bo Gao *et al.* [8] described how the semantic similarity can be measured based on the edge-counting and information content theory on Word Net. Implemented the similarity measure and strategies for semantic similarity measuring according to shortest path length. The results show that the new approach achieves high correlation value and distribution characteristics of correlation coefficient.

Verginica Barbu Mititelu *et al.* [9] discussed how to add derivational relations between literals and question answering task in Romanian Wordnet. Proposed two principles such as the Hierarchy Preservation Principle and the Conceptual Density Principle. Relevant answers for user query were found by calculating the similarity score between the words and the length of lexical chains between them is considered.

3. Methodology

In this information age, it is a deplorable state that despite the overload of information, we regularly fail to locate relevant information. This can be attributed to several factors, the most important being the absence of identification of context and semantics of the user query in fetching the required results. The aim of this paper is to generate concepts using Wordnet (Ontology) to improve the efficiency of information retrieval system.

The generation of Rough Ontology from Documents is highlighted in Figure-2. First a set of text documents is considered. The text in the documents is tokenized and stop words are removed. Then word stemming involves the removal of word suffixes leaving only stems or roots. Thus each document is transformed into list of terms along with their frequencies. The vocabulary of index terms for the corpus is formed by removing the most frequent and infrequent terms from the above input. The document term matrix is formed. Each row of the matrix document vector and each column represent the term and its prominence in various documents. K-means clustering algorithm is used to partition the terms associate with related concepts into k-equivalence classes. Now concepts and synonym set are assigned to each concept using word net.

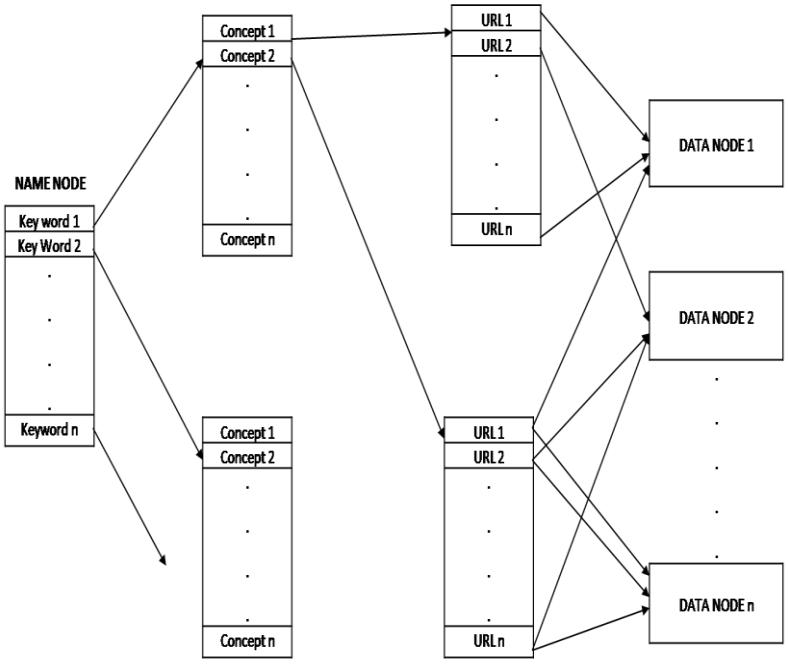


Figure 2. Mapping of Keywords with Concepts of Ontology

Retrieving the documents related to user query using Ontology is highlighted in Figure-3. The query passage is tokenized and stop words are removed. Words appearing in different morphological forms are mapped on to their common terms. Frequency of each term is calculated and least useful words are stripped. Synonyms are attached to each word present. The clustered query is mapped on to ontology constructed in phase-1 and relevant documents are ranked and retrieved. The documents are displayed to the user.

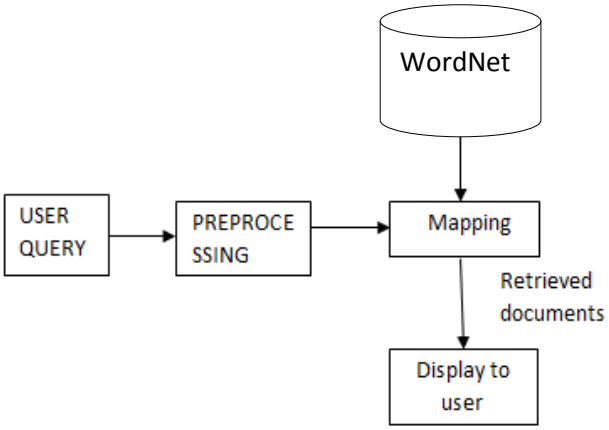


Figure 3. Mapping of Keywords with Concepts of Ontology

The detailed algorithms are presented in the following sections.

3.1. Pre-Processing

Documents containing repeated occurrences of keywords, not all the words notable for representing the semantics of a document are to undergo pre-processing in order to derive a well-defined index terms. In this process the documents are made to undergo a series of transformations like removing the Stop words, Stemming process and Stripping of least useful keywords. Stop words are the most common function words used in our day to day life such as the, is, at *etc.* Corpus words are tokenized and each word is checked if it is a stop word or not, if not then removed.

Stemming process is that in which related words are mapped to its stem, even if the stem is not itself a valid root. In general terms with a common stem own similar meaning. For example; the word "FISH" is found with different stems like:

- FISH
- FISHER
- FISHING
- FISHES

Frequently, the performance of an information retrieval system will be enhanced if term groups such as this are changed into a single term. Porter stem Algorithm is used based on the idea of suffixes.

The importance of the document to the concept associated with the keyword is determined by the frequency of keyword in a document. The keywords with least frequency as well as the keywords that occur very often in most of the documents are less likely to be significant in finding the weight of a document. To strip the least useful words we first calculate the corpus frequency *i.e.*, the number of documents containing the term using Hadoop map reduce and then those with least frequency and highest frequency are removed.

3.2. Clustering

We make over the pre-processed data into term-document matrices, each of which limited to the collection of documents belonging to a selected corpus. This can be used in grouping the similar terms. Clustering methods are used to identify the groups of terms based on their similarity estimates. We can use tf-idf estimate of a term in a document as a component of the term vector. Tf-idf estimates are given by:

$$\text{Tf-idf} = tf_{ij}/n*\log(m/df_i)$$

tf_{ij} is the term frequency of ith term in jth document
n is number of words in jth document
m is total number of documents
df_i number of documents in which ith term appears

A cluster is described as a set of similar objects or entities collected or grouped together. All entities within a cluster are alike and the entities in different clusters are not alike. Each entity may have multiple attributes, or features and the likeness of entities is measured based on the closeness of their features. Therefore, the crucial point is to define proximity and a method to measure it. To estimate the proximity of terms we use Euclidean distance. The distance between two things (a & b) using Euclidean distance is given by:

$$d(a,b) = SqRt\,[\sum_{i=1}^{n}(bi - ai)2\,]$$

For calculation of each of these term frequency, total number of words in a document, number of documents containing the term and finally tf-idf is done using Hadoop map reduce.

K-means is a common and well-known clustering algorithm. This algorithm starts with the selection of the k initial random cluster centers from the n objects. Each remaining object is assigned to one of the initial chosen centers based on similarity measure. When all the n objects are assigned, the new mean is calculated for each cluster. These two steps of assigning objects and calculating new cluster centers are repeated iteratively until the convergence criterion is met.

In this phase the query passage is tokenized and stop words are removed. Stemming is performed on these words and is mapped on to their common stems. This list of stems in the query passage obtained is mapped with reference to the k equivalence classes obtained during generation of ontology.

Now for a given keyword to search the file in which the keyword is most frequently occurred. Identify the documents which are having more relevance to the given user query. If multiple documents exist in this group rank them based on their similarity to the query and is represented in Table 1.

Table 1. Calculating Relevance Score for the Word to Retrieve Documents

Word	W_i				
File ID	F_{i1}	F_{i2}	F_{i3}	….	F_{iNi}
Relevance Score	0.00045	0.00123	0.000432	0.00743	0.0573

4. Experimentation

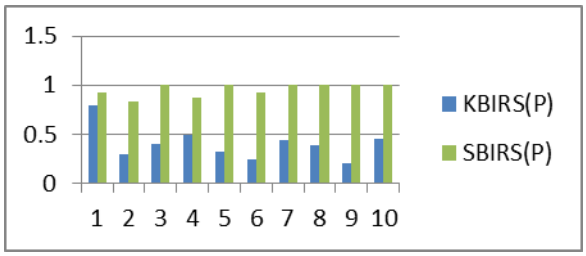
4.1. Environment Setup

The experiments were performed on a 4 node cluster equipped with Hadoop. This project has been provisioned with one Name Node and four Data Nodes. The Name Node was configured to use two 2.5-GHz CPUs, 2 GB of RAM, and 500 GB of storage space. Each Data Node was configured to use two 2.5-GHz CPUs, 2 GB of RAM, and 500 GB of disk storage. Besides this, all the computing nodes were connected by a gigabit switch. BOSS GNU Linux 4.1., Hadoop 0.20.1, and Java 1.6.0_6 were installed on both the Name Node and the Data Nodes. For cloud set up we used Xen hypervisor for virtualization and

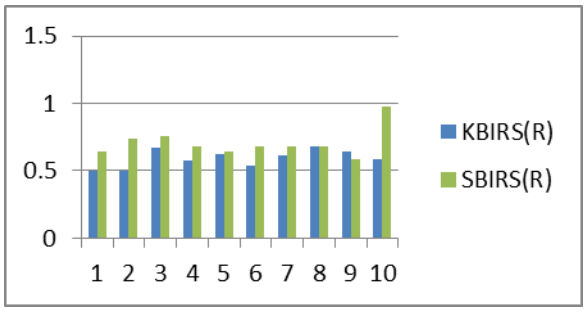
Eucalyptus for cloud infrastructure establishment. For performance monitoring of hadoop cluster, ganglia-3.6.0.is used.

The evaluation of IR model is carried out by using the objective retrieval quality testing methodologies. The documents retrieved for the given user preference keyword are evaluated by using metrics such as Precision (P), Recall (R), F-measure (F), Error Rate (E), Accuracy (A).

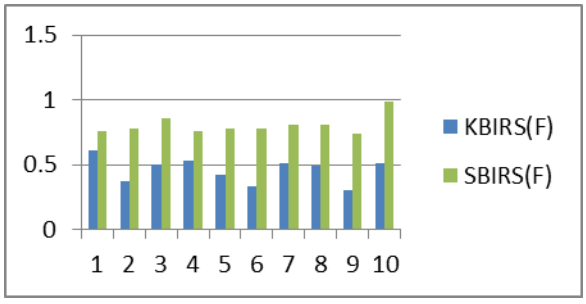
For experimentation we have taken 400,800 and 1000 documents from corpus data set and the observed results are tabulated.



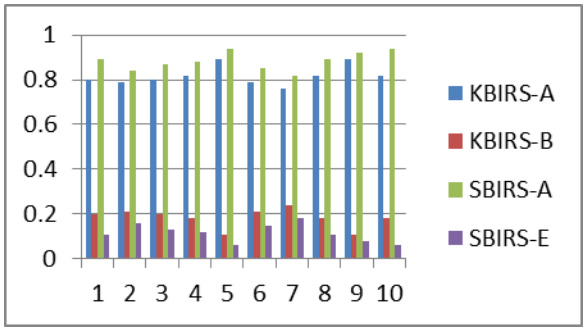
Graph 1. Precision versus Number of Documents in HDFS



Graph 2. Recall versus Number of Documents in HDFS



Graph 3. F-Measure versus Number of Documents in HDFS



Graph 4. Accuracy versus Error Rare for Keyword and Semantic Based Model

5. Conclusions and Future Work

In this paper, "Secure semantic Crypto-Information retrieval model based on MapReduce using k-means algorithm" is developed. The developed model is compared to the existing model where key based information retrieval system (KBIRS). The Hadoop cluster framework is developed keeping in view of processing large volumes of data securely. The developed framework helps to analyze more data intensive applications in a better contrast than that of the existing models.

The proposed method of Secure Semantic Symmetric Encryption based Ranking of Encrypted data in the extended Hadoop framework can be further improved for its increased applicability. The following issues may be considered for prospective development in this area.

Concept formation can be evaluated with the exploration of other clustering algorithms like hierarchical and fuzzy c means clusters with MapReduce. The framework can be deployed in a cloud environment to provide computation as a service on other real or benchmarked data sets.

It is also better to provide service as an object in which user facial features or other biometrics can be taken once again before providing service to user, to avoid service misuse among authorized users.

The Information Retrieval application deployed in the extended framework can also be used to support multiple keywords search with negation words. New approaches can also be designed for IDF factor to preserve the order while calculating the score for the keywords provided by the user.

The above issues may be considered for future task in the overall development of SSSE based ranking encrypted data systems.

References

[1] P. U. Bulakh, "Application of semantic similarity using ontology for document comparison", IJRCM, vol. 3, no.12, **(2013)**.
[2] T. C. Jepsen, "Just What Is an Ontology, Anyway"? IT Professional, vol. 11, no. 5, **(2009)**, pp. 22-27.
[3] M. Obitko, V. Snasel and J. Smid, "Ontology Design with Formal Analysis", Communications In computing, **(2014)**, pp. 302-310.
[4] A. Amine, "Evaluation of Text Clustering Methods Using WordNet", The International Arab Journal of Information Technology, vol. 7, no. 4, **(2010)**.
[5] M. Zurini, "Word Sense Disambiguation using Aggregated Similarity based on WordNet Graph Representation", Informatica Economică, vol. 17, no. 3, **(2013)**.
[6] C. Bouras and V. Tsogkas, "A clustering technique for news articles using WordNet", Elsevier, vol. 36, **(2012)**, pp. 115-128.
[7] K. T. Reddy, M. Shashi and L. P. Reddy, (2008), "Hybrid Clustering Approach for Term Partitioning in Document Data Sets", Artificial Intelligence and Pattern Recognition, **(2007)**, pp. 165-172.
[8] J. B. Gao, B. W. Zhang and X. H. Chen, "A WordNet-based semantic similarity measurement combining edge- counting and information content theory", Elsevier, vol. 39, **(2014)**, pp. 80-88.
[9] V. B. Mititelu, "Increasing the Effectiveness of the Romanian Wordnet in NLP Applications", Computer Science Journal of Moldova, vol. 21, no. 3, **(2013)**.
[10] A. M. A. Zoghby, A. S. E. Ahmed and T. T. Hamza, "Arabic Semantic Web Applications – A Survey", Journal Of Emerging Technologies In Web Intelligence, vol. 5, no. 1, **(2013)**.
[11] L. Meng and J. Gu, "A New Method for Calculating Word Sense Similarity in WordNet", International Journal of Signal Processing, vol. 5, no. 3, **(2012)**.
[12] S. Vijay, "A Comparison of Different Measures to Evaluate the Semantic Relatedness of Text and its Application", IJRTE, vol. 1, no. 1, **(2012)**.
[13] Menaka S. and Radha N., "Text Classification using Keyword Extraction Technique", IJARCSSE, vol. 3, no. 12, **(2013)**.

Authors

P. Srinivasa Rao, he is currently working as Associate Professor in CSE Department of MVGR College of Engineering. He is having Over 10 years of teaching experience. His research includes Data warehousing and Mining, Distributed Computing, Image Processing *etc.*

K. Thammi Reddy, he is the Director of Internal Quality Control (IQC) and Professor of CSE at Gandhi Institute of Technology (GITAM).He is having Over 20 years of experience in teaching, Research, Curriculam Design and consultancy. His research areas include Data warehousing and Mining, Distributed computing, Network Security *etc.*

MHM Krishna Prasad, he is the Professor of CSE at JNTUK Kakinada, He is having Over 24 years of experience In teaching, Research, Curriculum Design and consultancy. His research areas include Data warehousing and Mining, Distributed computing, Computer Networks *etc.*