ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 3, March 2016

Swarm Search Using WordNet and Hadoop

Prof. Avinash Palave[1], Archana Thakur[2], Priyanka Ranpise[2], Ragini Katta[2], Aasma Kazi[2]

Assistant Professor, Dept. of Computer, TCOER, Savitribai Phule Pune University, Pune, India[1]

B.E. Student, Dept. of Computer, TCOER, Savitribai Phule Pune University, Pune, India[2]

ABSTRACT:Now a days handing of big data is not easy because its size and complexity .The capability of removing or take out useful information from these large datasets of data, because of its volume, variability, and velocity is nothing but the big data , it was impossible earlier to do it. PSO is a naturally distributed algorithm Particle Swarm Optimizers are naturally distributed algorithms in that solution to problem is form by interaction between different particles. This is concept related to Data mining. It includes, Particle Swarm Data Mining Algorithms in which we implemented and tested across a natural Algorithm and a Decision Tree Algorithm . From the archived results, Particle Swarm Optimizers proven that it is to be a sufficient for classification tasks. The data which used for experimental testing are commonly existing standard for rule discovery algorithms reliability ranking .Also the feature selection algorithm used to remove a redundancy in document and gives most relevant document .Wordnet provide you different synonyms for search the given word in hadoop document.

KEYWORDS: Particle swarm optimization algorithm; Feature selection; Hadoop;Mining Big data stream; Decision tree

I. INTRODUCTION

Information Retrieval is a process of finding the documents in a collection based on a specific topic .The information which is want or need by user is show as query. Document which satisfy for given query this document is called as relevant document. The documents which not safety the given query are said to be non-relevant. An IR may use the query to classify the documents in a collection, returning to the users subset of documents that satisfy some classification criterion.

There are many search software's are available to search a documents from high dimensional and text form. The information retrieve from bible corpus is big challenges . Sometimes the relevant documents may not contain the specified keyword. The absence of the given term in a document does not necessarily mean that the document is not a relevant. Because more than one terms can be semantically similar although they are alphabetically different. In old days Decision based tree are used to data mining but now a days PSO algorithms are having greater performance . The basic three Algorithms are used for Information Retrieval in Hadoop is Particle Swarm search ,feature selection and the apache lucence. Apache lucene is for indexing the document to provide better data mining.

II. LITERATURE SURVEY

Big data is term related to high dimensional datasets which having large size and complexity[8]. This datasets are not easy to handle with old technology .Hence Big data mining is introduce to over come the challenges of data mining . Data mining having capability to fetch the useful data from datasets . Velocity, variety, volume are the 3V challenges which canbe handle by data mining[8]. Decision tree can be used represent decision and also decision making in data mining. Decision tree illustrate data but not decisions. PLANET[5][1] is a framework used to learn tree models by using large datasets. PLANET utilizes Map Reduce functionality to provide scalability. Particle swarm optimization (PSO) is algorithm uses the concept of swarm search[4]. PSO algorithm is used for better efficiency of data. PSO is also use as a tool for data mining. Particle Swarm optimization Algorithms are competitive, not only with other evolutionary techniques, but also with industry standard algorithms. We also implement swarm intelligence for data mining[7]. PSO algorithm is initialized with population of random solutions[7]. In this we implement PSO as well as Feature selection algorithm for better performance. Feature selection algorithm check redundancy in datasets for indexing the document. Hadoop is framework we implement to store the database[9]. Hadoop is used to process on huge amount of data sets on system clusters develop from product hardware.

ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 3, March 2016

III. EXISTING SYSTEM

In Existing system, we find the documents based on complete keyword given in search box. This method misses some important documents.

A. *Decision tree based:*

Decision tree learning uses, decision tree as a predictive model which maps observation about an item to conclusions about the items destination value [1].It is one of the predictive modeling approach used in data mining machine learning and statistics. Decision tree can be used explicitly represent decision and decision making in data mining, decision tree illustrate data but not decisions.

A decision tree consists of nodes ,branches and leaves ,a node consists of query about value of an attribute. Branch is a connection between nodes ,that is established based on answer of the corresponding query. Leaf is the end point in the tree. Here, Decision Tree implementation exploits recent research decreasing the computational complexity of decision tree assessment, allowing linear scalability with data amount and number of nodes. This algorithm processes data in large amount, allowing scaling unconstrained by combined cluster memory. The implementation base both classification as well as regression and is completely integrated with the R statistical language and the rest of our advancing analytics and machine learning algorithms, as well as our interactive Decision Tree visualizer. There are some efforts that improve performance of the decision tree. When processing of high amount of Big data by paralyzing inductive process in distributed environment . PLANET [5] is framework for learning model .It utilize MapReduce to provide Scalability .

IV. PROPOSED ALGORITHM

A. *Particle Swarm Optimization:*

Particle Swarm Optimization (PSO) is a universal optimization algorithm for dealing with problems in which a best solution can be expressed as a point in an n-dimensional space. PSO shares many similarities with transformative computation techniques such as Historical Algorithm. Over a number of iterations ,group of variables have their values adjusted adjacent to the member whose value is closest to target at any given moment[4][2].

B .Feature Selection Algorithm:

Feature selection also known as variable selection, virtue selection or variable subset selection. Feature selection is the process of choosing a subset of related features for use in model construction .It can be seen as the combination of search technique for proposing new feature subset with evolution measure which count the different feature subsets. Feature selection algorithm is to check each possible subset of feature finding the one of which minimize the error rate[6][2].It has been an active and also beneficial field of research area in pattern approval , machine learning and data mining .this also implement fin theory and practice to be enhancing efficiency increasing predictive accuracy and decreasing complexity of learn result [6][2].

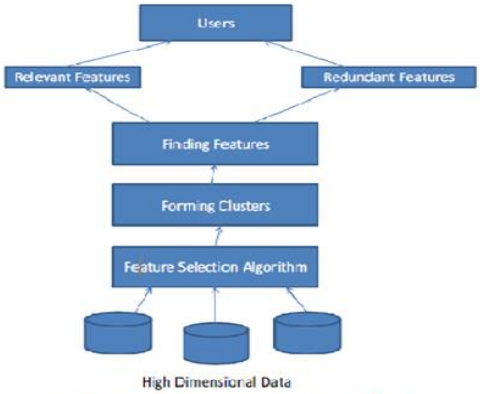

Figure 1 Flow Chart for Feature Selection

ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 3, March 2016

Description of Figure as follows:

1. Collect a training data set from the specific domain.
2. Shuffle the data set.
3. Break it into *P* partitions, (say *P = 20*)
4. For each partition (*i = 0, 1, ..., P-1*)
    a. Let *OuterTrainset(i)* = all partitions except *i.*
    b. Let *OuterTestset(i)* = the *i*'th partition
    c. Let *InnerTrain(i)* = randomly chosen 70% of the *OuterTrainset(i).*
    d. Let *InnerTest(i)* = the remaining 30% of the *OuterTrainset(i).*
    e. For *j = 0, 1, ..., m*

Search for the best feature set with *j* components, *fsij*.using
    leave-one-out on *InnerTrain(i)*
Let *InnerTestScore*ij = RMS score of *fs*ij on *InnerTest(i).*
    End loop of *(j).*

    f. Select the *fsij*with the best inner test score.
    g. Let *OuterScorei*= RMS score of the selected feature set on *OuterTestset(i)*
    End of loop of *(i).*

5. Return the mean Outer Score.

V. PSEUDO CODE

Pseudo Code for Particle Swarm Optimization algorithm is as follows:

Step 1: For each particle calculate fitness value.
Step 2: If the fitness value is better than the best fitness value (pBest) in history.
Step 3: Set current value as the new pBest.
Step 4: go to 8.
Step 5: Choose the particle with the best fitness value of all the particles as the gBest.
Step 6: Else For each particle calculate particle velocity according equation (a).
Step 7: Update particle position according equation (b).
Step 8: End.

VI. SYSTEM ARCHITECTURE

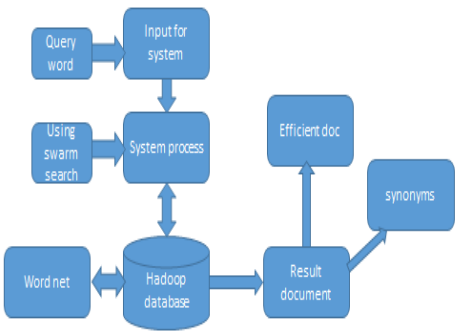

Fig. System Architecture

ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 3, March 2016

The description of system architecture is as follows:

Query Word:

Query Word is Word or Document which we searching from existing set of document or our Database. It is input to system.

Swarm Search:

The input which is given to system that process by using PSO algorithm. Stemming is done in on data which stored on Hadoop for indexing. In swarm search the query word that is to be search is process by swarm algorithm and gives the result document.

Input for System:

Query word or document which is to be searched from give datasets is the input for system. Also the document from which we search is input to system.

System Process:

System process including some process and algorithms .The documents which stored on hadoop is indexed by content wise and also removing the stop words in documents for efficient search . Word is process by PSO n feature selection algorithm and find out the synonyms of the word and we get the result document whose wait is greater.

WordNet:

Wordnet is large database which contain data of English , nouns , Verbs , adjectives and adverbs are grouped into sets of related synonyms . Wordnet superficially resembles the sources, in that if groups words together depend on their meaning . However, there are some important distinctions .The main relation among words the in Wordnet is synonymy ,Synonyms is word that denotes the same concepts and are interchangeable in many context and grouped in the unordered sets.

HadoopDatabase :

In hadoop database we store all files. HDFs ishadoop data stores large files across multiple computers. Which stores the data in range of gigabytes to terabytes. An advantage of using HDFS is data consciousness between the job tracker and task tracker that is in Slave mode.

VII. TECHNOLOGY AND CONCEPTS

A. *Mining Big Data Streams:*

Big data usually includes data sets having sizes beyond the capacity of commonly used software tools to capture, data curation, managing and processing data within a passable elapsed time. Big data size is a regularly moving target, as of 2012 ranging from a few dozen of terabytes to many petabytes of data. It is the set of methods and technologies that require new form of integration to uncover large hidden values form large datasets that are distinct , complex and of a massive scale. META GROUP analyst Doug Laney defined data growth difficulties and opportunities as being three-dimensional, that is increasing volume that is amount of data, velocity that is speed of data in and out and having different variety that is range of data types and sources . Now much of the industries are using this "3Vs" model for characterized big data. In 2012, META Group updated its definitionand it can be defined as follows: Big data is the data having high volume, high velocity, or high variety information assets that require new forms of processing to enable enlarged decision making, understanding discovery and process optimization . Big Data represents the information that can be characterized by parameter such as High Volume, Velocity and Variety to require desired Technology and Analytical techniques are used for its transformation into Value[7][2]. The 3Vs have been extended to other complementary features of big data :

- Volume: Big data does not sample. It just examines and tracks what happens.
- Velocity: The big data is available in real-time usage .

ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 3, March 2016

- Variety: Big data draws from different types of data such as text, photos, voice recording, video; plus it completes missing segments through data fusion.
- Machine Learning: big data generally doesn't ask why and simply find patterns.
- Digital Footprint: The big data is often a cost-free by product of digital interaction.

B. *Hadoop:*

Apache Hadoop[9] is an open source software framework which is written in java language for distributed storage and processing of huge amount of data sets on system clusters develop from product hardware. All the programs in Hadoop are designed with a primitive assumption that failures of hardware or individual machines are standard and thus must be automatically handled in software by framework. The core of Apache Hadoop consist of storage called as Hadoop Distributed File System and MapReduce is a processing part that .Hadoop splits files into large data sets called blocks and distributes it over the nodes of cluster. To perform the operations on data, HadoopMapReduce is process data and transfer package for nodes to process in parallel, depend on the data each node require to process. This way takes benefits of data locality nodes manipulating the data due to this data to be processed faster and more efficiently than would be in a more ordinary super computers that depend on a parallel file system where computation and data are connected through high-speed networking.

The Hadoop framework is mostly written in the programming language that is Java, with some basic code in C and command line services written as Shell script. For the end-users, though MapReduce Java code is mostly used,"Hadoop Streaming" can be implemented with any programming language. Hadoop Streaming is used to implement "map" and "reduce" are the parts of the user's program. Other similar projects expose other higher-level user interfaces.

VIII. RESULT AND DISCUSSION

In this we use Data mining and Hadoop technology to improve performance of existing system. Hadoopis framework to store high dimensional data in huge amount. This paper is based on information retrival system based on Hadoop. In our system when we need some file or any document there is no need to remember particular file name just know about any word in that file. This word is query word for our system. Which is search by using two algorithms PSO and Feature selection also for efficient data mining we implement apache lucene algorithm.

After processing query word we get result document. Word net dictionary provide synonyms to query to find result document.

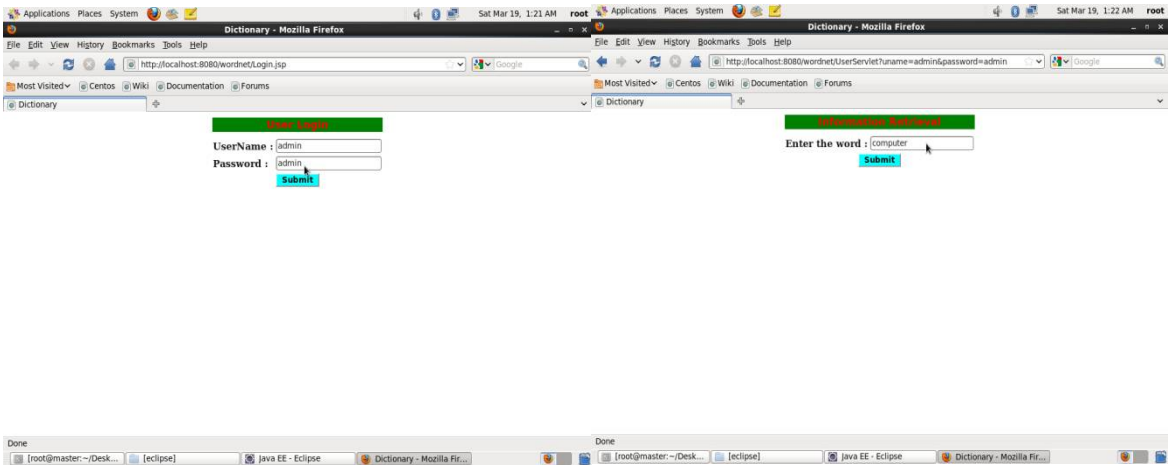

Fig.1 Step 1: Enter the User Name and Password to Log in

Fig.2 Step 2: Enter the query word which is to be searched

ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 3, March 2016

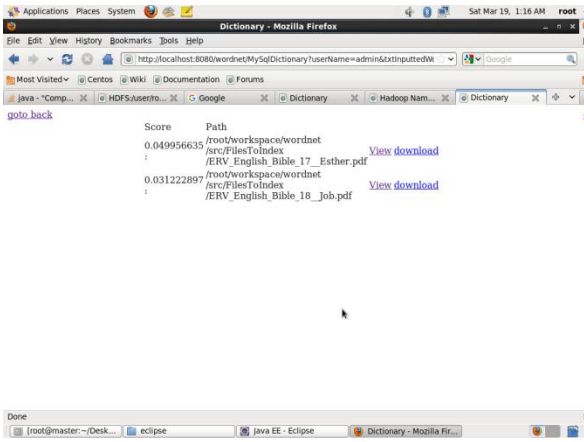

Fig.3 Step 3: Display list of Paths of result file with view and download option

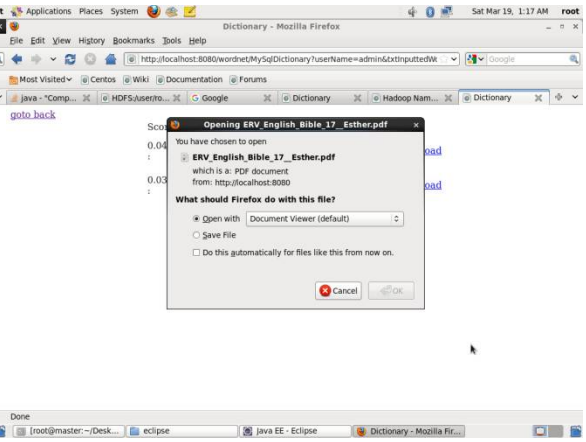

Fig. 4 Step 4: After choosing view option to open result document

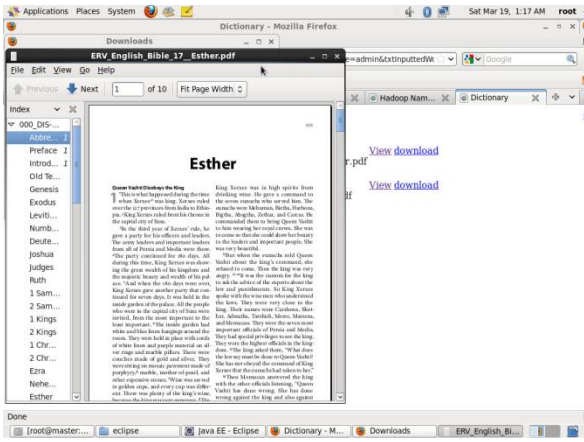

Fig. 5 Step 5: Open the result document

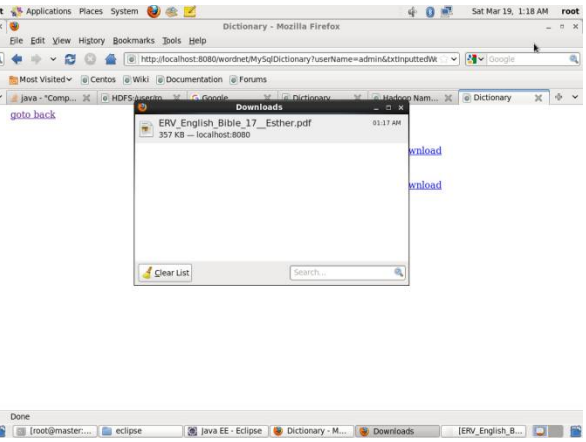

Fig. 6 Steps 6: Download the result document

IX. CONCLUSION

In this paper by using particle swarm optimization and feature selection we implement the concept of data mining to increase our computational speed and accuracy of search document.

By using wordnet synonyms are find out for query word .Here we implementing system in which indexing is done by apache lucene for fast searching.

REFERENCES

1. ArintoMurdopo, "Distributed Decision Tree Learning for Mining Big Data Streams", Master of Science Thesis, European Master in Distribut-ed Computing, July 2013.
2. Simon Fong, Raymond Wong, and Athanasios V. Vasilakos, "Accelerated PSO Swarm Search Feature Selection for Data Stream Mining Big Data",IEEE, DOI 10.1109/TSC.2015.2439695,2015.
3. Simon Fong, Suash Deb, Xin-She Yang, Jinyan Li, "Metaheuristic Swarm Search for Feature Selection in Life Science Classification", IEEE IT Professional Magazine, Volume 16, Issue 4, pp.24-29August 2014.
4. Tiago Sousa ,Arlindo Silva , Ana Neves ,"Particle Swarm based Data Mining Algorithms for classification tasks",ELSEVIER,Parallel Computing 30 (2004) 767–783,May 2004.
5. B. Panda, J. S. Herbach, S. Basu, R. J.Bayardo,"PLANET: Massively Parallel Learningof Tree Ensembles with MapReduce",VLDB 2009, pages1426{1437, Lyon, France, 2009.
6. M. Ramaswami and R. Bhaskaran,"A Study on Feature Selection Techniques inEducational Data Mining",JOURNAL OF COMPUTING, VOLUME 1, ISSUE 1, ISSN: 2151-9617, DECEMBER 2009.
7. CrinaGrosan, Ajith Abraham and Monica Chis, "Swarm Intelligence in Data Mining", Springer,Studies in Computational Intelligence (SCI) 34, 1–20 (2006).
8. Wei Fan, Albert Bifet "Mining Big Data: Current Status, and Forecast to the Future",SIGKDD Explorations, Volume 14, Issue 2.
9. http://hadoop.apache.org/