© 2016 IJSRSET | Volume 2 | Issue 1 | Print ISSN : 2395-1990 | Online ISSN : 2394-4099
Themed Section: Engineering and Technology

Hadoop based Information Extract from Text Document

Deepak Motwani, V. K. Chaubey, A. S. Saxena

Department of Computer Science and Engineering,. Mewar University, Rajasthan, India

ABSTRACT

Hadoop is one of the generally received bunch figuring structures for handling of the Big Data. Despite the fact that Hadoop seemingly has turned into the standard answer for overseeing Big Data, it is not free from constraints. In nowadays developing technology researchers, students prefer all documents in txt format and doc format. Most text files are available in pdf format as per demand. Even all research papers are available in pdf format only and extracting a text from pdf format is one of the most difficult jobs. So for text extraction from multiple pdf files we have to apply some algorithms so that text extraction process takes place in comfortable mode. Text extraction is the basic step which we bear to follow before making a motion for further processing. We begin with the concise discussion concerning to the keyword. Steps involved in text extraction from any txt file. In this paper, we use a keyword based extraction method for extracting the text from txt file and with the help of these keywords we can get all the detail on that part of the research paper or any pdf file. Here we are also using the multithreading approach. Our approach is able to extract text in very less time, so time complexity is very less. The aim of this paper is to extract the text on the basis of particular keyword which is useful for the new researcher.

Keywords: Hadoop Big Data, Text Extraction, Keyword Based Extraction, Map Reduce.

I. INTRODUCTION

The 'Hadoop HDFS' is a distributed file system, which is used to store datasets among all nodes in a cluster. This is absolutely necessary when trying to run a MapReduce program. The design of HDFS was meant to perform on commodity hardware, having much higher fault rate. compared to expensive server systems At that place are yards of research papers, books are published each day hence it gets really difficult to read all the books and papers to ascertain the material of our purpose. To get the best from this difficulty researchers has given much text extraction method so that we can able to extract needed data regarding the topic we want. In this report, we are working on keyword based text extraction method. The keyword extraction methods are essentially split into three categories first one is linguistic method second is statistical method and the final one is mixed methods. The main focus of Statistical methods is on non-linguistic features of the text like position of the keyword, inverse file frequency and so on. In statistical method computation is really less and real gentle to use and able to yield best results. As we totally know that Keyword plays a decisive role in the extraction of valuable information which is being read by the user. Then on that point are some effective keywords which are necessary for the text extraction. Keyword just contains only 7 letters in itself, but this small unit done a large job in conveying the meanings of whole txt file.Thus many other applications like text Summarization, clustering, data retrieval can take the advantage of it [1] [3].

By keyword based extraction we can able to identify the most relevant information with the help of keywords from pfd files [2][4].. We have experienced in many articles there are some methods which recompense the attention on linguistic features like syntactic structure, part of speech and semantic qualities are likely to total value, functioning sometimes as filters for bad keywords [6]. Many of the linguistic method are named as mixed

methods Some of the linguistic methods are in fact mixed methods, they are integrated with some common statistical incorporating some linguistic methods with common statistical procedures like inverse document frequency and term frequency [8]. In these days most of the documents are electronically available. There is a Domain-independent keyword extraction technique is available which doesn't need large corpus, it has many applications. Considering an object lesson if we are looking for new web first we are in a haste to know the master content of that page by anyhow or any men like highlighting of special keyword for easy descent. If a researcher is interested to know the actual summary of the theme or the main office of the paper, for that they need some keywords. Here the extraction of keyword done without using corpus is very helpful. Many times we use the word count for document overview. [10]; however, a further powerful tool is enviable.

II. METHODS AND MATERIAL

A. Related Work

Shernandez, ezpeleta provides a technique to implement self tuning in Big Data Analytic systems. Hadoop's performance out of the box leaves much to be desired, leading to suboptimal use of resource, time and money. To calculate each word in corpus of all txt files Markov Chain is being used [5]. The technique explained Markov Chain for pdf document d and term t having two states C,T where the transition probability from C to T is define as the probability of the given term being observed in document d from all pdf files. The probability of T to C movement is define as the probability of the term being observed out of all values in d. Here the writer remarks that if the words are raised at the same state having the same frequency of all other words, in the pdf document than they need a very less description. (Named the background distribution), the simple description of the text file is required only when if the document diverges most form background distribution. The technique is being practiced to match and often beat, TF-IDF in provisions of accuracy when tested over a corpus of document abstract from IEEE, ACM [5].

One other method for keyword extraction is explained by Oshawa, et al. [7], which gives the answer to the problem in a different manner. As most of the keyword based extraction methods are depended on statistical information collected from frequency occurrence in the pdf file and the method is known as key graph which works to cluster the associated items in different groups for determining which word of the document is working as a representative of the content available in the document. The graphical representation of the pdf file is being made with the help of nodes and edges using a key graph method where nodes represent the usual occurrence of co-occurrences within the text file. Word cluster is being placed by putting maximum connected sub graphs in making a graphical record of the document. The graph created with nodes that are having edges between two different clusters is the way to identify the candidate keywords. Then, with the help of these keywords are the results of the author's concept, thinking, and methods to answer the question of how to select the relevant document or how to get the information of any relevant document especially in the case of researchers. Then the ranking of these candidate keywords is being done on the basis of the probability for every cluster which they join. Here to join word represents the joining of two clusters. (Efficiently, the most universal word used to link up these clusters). When the testing is performed on the key graph than it is observed that it can be able to surpass and match, TF-IDF in chain of test is being feed by its developer [7]. In addition, on social media data which is being collected in year 2008 chain of test is being run, presidential elections prove that the key graph is being capable to identify keywords in a noisy environment with a vast quantity of extraneous information [9].

B. Methodology

Apache Hadoop is an open source software system for capacity and substantial scale preparing of data sets on bunches of item equipment.

- ✓ Hadoop MapReduce: a software model for large scale information processing.
- ✓ Data node—The data nodes are the repositories for the data, and consist of multiple smaller database infrastructures that are horizontally scaled across compute and storage resources through the substructure.
- ✓ Larger big data repositories will have numerous data nodes. The critical architectural concern is that unlike traditional database infrastructure, these data

nodes have no necessary requirement for locality of clients, analytics, or other commercial enterprise intelligence.

- ✓ Client—The client represents the user interface to the big data implementation and query engine. The client could be a server or PC with a traditional user interface.
- ✓ Name node—The node name is the equivalent of the address router for the big data implementation. This node maintains the index and placement of every data node.
- ✓ Job tracker—The job tracker represents the software job tracking mechanism to pass out and aggregate search queries across multiple nodes for ultimate client analysis
- ✓ Variety—Extends beyond structured data and includes semi-structure or unstructured data of all sorts, such as text, sound recording, video, click streams, log files, and more.
- ✓ Volume—Comes in one size: large. Organizations are awash with data, easily amassing hundreds of terabytes and petabytes of data.
- ✓ Velocity—Sometimes must be analyzed in real time as it is streamed to an organization to maximize the data's business value.

The distribution of work to Task Trackers is exceptionally basic. Each Task Tracker has various accessible spaces, (for example, "4 openings"). Each dynamic Map or lessen assignment takes up one space. The Job Tracker dispenses work to the tracker closest to the data with an approachable distance. On that point is no thought of the present framework heap of the assigned machine, and henceforth its genuine accessibility. In the case that one Task Tracker is moderate, it can submit the whole Map Reduce work—especially towards the close of an occupation, where everything can wind up sitting tight for the slowest errand. With theoretical execution empowered, be that as it may, a solitary errand can be performed on different slave node[11]

- Clients (one or more) submit their work to Hadoop System.
- When Hadoop System receives a Client Request, first it is received by a Master Node.
- Master Node's MapReduce component "Job Tracker" is responsible for receiving Client Work and divides into manageable independent Tasks and assign them to Task Trackers.
- Slave Node's MapReduce component "Task Tracker" receives those Tasks from "Job Tracker" and perform those tasks by using MapReduce components.
- Once all Task Trackers finished their job, Job Tracker takes those results and merges them into the final solution.
- Finally Hadoop System will send that final result to the Client[12].

III. RESULTS AND DISCUSSION

The step by step procedure for extracting the text from any pdf file on the basis of keyword is explained. In the market lots of systems are available for text extraction. But the motto of the paper is to make the researchers to make the more advance system for the process by making new steps for text extraction. Our system results in a smooth and less time consuming text extraction systems. In this step when fetched data from word document select the paper title, author, abstract and select browse file it shows the execution time in 38144 milliseconds which is far better from older.

For performance evaluation, we considered Hadoop three nodes cluster with homogeneous hardware property ,each node in cluster has a 3.9 GB Ram,Intel-I3 cpu @2.20 GHz. We setup a cluster on Red hat 5 Linux with hadoop 1.2.1 stable reslease,used JDK 1.7 and ssh configuration with 1 name node and 3 data node for experiment configured files such as maped-site.xml,core-site.xml,hdfs-site.xml and setup default values with replication factor 2 and block size 64 MB.We used pdf files data 2.0 GB for our analysis which has better result in overall hadoop performance. Effected by reusing JVM, increases the number of mappers and reducers, input reduce buffer parameters, sort buffer size sort factors, compressing Mappers output.

Max. no. of mappers slots= (CPU cores – 1 reserved cores for Hadoop daemons) * 0.95 to 1.75(0.95 to 1.75 is CPU hyper threading factor)

Cluster mappers capacity= max number of mappers slots*number of nodes[13]

Table 6.4 hadoop performance: Effected by reusing JVM, increases the number of Mappers and reducers, input reduce buffer parameters, sort buffer size sort factors, compressing Mappers output.

Table 4 Hadoop Performance

Method	Map Time(s)	Reduce Time(s)	Total Cpu Time(s)
Base line system	263	10	273
Effected by resuming JVM	217	10	228
Mapper and reducer slots	202	10	212
Input Reduce buffer	191	5	196
Sort buffer size and sort factor	174	8	182
Compressing mapper output	215	34	251

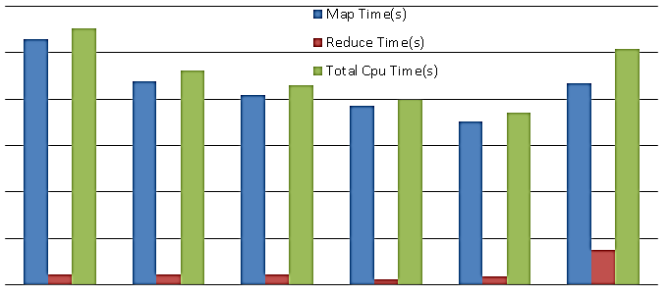

Figure 1 : chart for Map time, Reduce time, CPU time of the and proposed method using text data

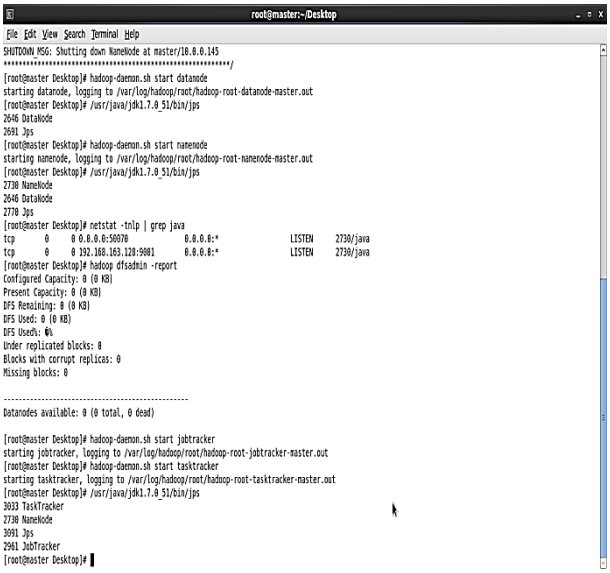

Figure 2 : JSP command to check NameNode, TaskTracker, JobTracker

IV. CONCLUSION

Limited database usefulness is not by any means the only reason Hadoop hasn't assumed control over the world. To start with, open-source software is notable for being exceptionally variable in character, with some of it being extinct and out unusable. The step by step procedure for extracting the text from any text file on the basis of keyword is explained in this paper. In the market lots of systems are available for text extraction. But the motto of the paper is to make the researchers to make the more advance system for the process by making new steps for text extraction. Our system results in a smooth and less time consuming text extraction systems.

V. REFERENCES

[1] M. Andrade and A. Valencia, Automatic extraction of keywords from scientific text: application to the knowledge domain of protein families, Bioinformatics, Vol. 14(7), , pp. 600-607, 1998.

[2] Y. Matsuo and M. Ishizuka, ”Keyword extraction from a single document using word co-occurrence statistical information,” International Journal on Artificial Intelligence Tools, vol. 13, no. 1, pp. 157-169,
2004.

[3] L. Plas, V. Pallotta, M. Rajman, and H. Ghorbel. Automatic keyword extraction from spoken text. A comparison of two lexical resources: the EDR and WordNet. In Proceedings of the LREC,2004.

[4] Y. HaCohen-Kerner, ”Automatic extraction of keywords from abstracts,” in Proc. 7th Int. Conf. Knowledge-Based Intell. Inf. Eng. Syst., vol. 2773, pp. 843-849., 2003

[5] Christian Wartena, Rogier Brussee, and Wout Slakhorst. Keyword ex-traction using word co-occurrence. In Proceedings of the 2010 Workshops on Database and Expert Systems Applications, DEXA ’10, , Washington, DC, USA, 2010. IEEE Computer Society, pages 54–58,2010.

[6] A. Hulth.. Improved automatic keyword extraction given more linguistic knowledge. In Proceedings of EMNLP, pp 216-223, 2003

[7] Yukio Ohsawa, Nels E. Benson, and Masahiko Yachida. Keygraph: Au-tomatic indexing by co-occurrence graph based on building construc tion metaphor. In Proceedings of the Advances in Digital Libraries Conference, ADL '98, pages 12–, Washington, DC, USA, 1998.

[8] A Hulth, Combining machine learning and natural language processing for automatic keyword extraction. Stockholm University, Faculty of Social Sciences, Department of Computer and Systems Sciences (together with KTH). 2004,

[9] H. Sayyadi, M. Hurst, and A. Maykov. Event detection and tracking in social streams. In Proceedings of International Conference on Weblogs and Social Media (ICWSM), 2009.

[10] H. P. Luhn, A statistical approach to mechanized encoding and searching of literary information, IBM Journal of Research and Development, Vol. 1(4), , Pp.309-317, 1957.

[11] http://hadoop.apache.org/mapreduce

[12] F. N. Afrati and J. D. Ullman. Optimizing Joins in a Map-Reduce Environment. In EDBT, pages 99–110, 2010.

[13] Dr. Siddaraju, Sowmya C L, Rashmi K, Rahul M "Efficient Analysis of Big `Data Using Map Reduce Framework" International Journal of Recent Development in Engineering and Technology (ISSN 2347-6435(Online) Volume 2, Issue 6, June 2014