An Effective Text Processing Approach With MapReduce

Jigyasa Nigam , Sandeep Sahu

Abstract— **Information Extraction is a technology that is innovative from the user's point of view in the current information-driven world. Rather than indicating which documents need to be read by a user, it extracts pieces of information that are salient to the user's needs. Links between the extracted information and the original documents are maintained to allow the user to reference context for example Named Entity Recognition(NER). It helps machine to recognize proper nouns (entities) in text and associating them with the appropriate types. Common types in NER systems are location, person name, date, address, etc. There are several NER systems in the world. Such as GATE, CRFClassifier, OpenNLP and Stanford NLP(Natural Language Processing). The NER system works fast for limited amount of documents but drawback of this system is that it works slows for huge/large amount of data. To overcome the drawback of NER system, this paper, report the development of a NER which is based on Map Reduce, a distributed programming model. This development helps to achieve the fast extraction with better performance.**

Index Terms— **Big textual data, Distributed computing, Hadoop, MapReduce, Maxent Tagger, Named Entity Recognition (NER) , Natural Language Processing (NLP).**

I. INTRODUCTION

Information extraction is the process in which extracting data from Unstructured Data, semi-structured Data and structured Data. Unstructured Data does not have organized any pre-defined manner or model. Structured data have organize in any pre-define manner or model. Generally extract information of human language texts by uses of natural language processing (NLP). [9][7] Recent activities in multimedia document processing like automatic annotation and content extraction out of images/audio/video could be seen as information extraction. Information Extraction is a technology that is originates from the user's point of view in the current information existing world. Rather than indicating which documents need to be read by a user, it extracts pieces of information that are relevant to the user's needs. Links between the extracted information and the original documents are maintained to allow the user to reference context. Information is in different shapes and sizes. One important form is structured data, where there is a regular and predictable organization of entities and relationships.

In information extraction NER system play a role is more and more important. It easily recognizes such as persons and organizations can be extracted with reliability. But the problem is that when we use huge amount of data then its processing speed vary slow. Its improve time complexity and space complexity also. [3] [9]

So improve to speed and reduce to space proposed work in this paper to apply NER system with Distributed MapReduce framework. Using the MapReduce framework with NER system got fast information extraction and reduced copy with accuracy.

In this paper use one of the NER system Stanford-POStagger(Part of speech tagger) [7] in which Maxent Tagger model to use extract the information in the form of Name Entity recognition.

The paper is organized as follows. Section 2 introduces related work. We report on
the design of the proposed distributed text parsing system in Section 3. Finally, in Section 4, we give the conclusion.

II. Related Work

In this section, introduce the Stanford-POStagger parser[7] , MapReduce programming model[5] and Hadoop[5]. These are used by the proposed system in single system and distributed environments. First, the Stanford parser, proposed by the NLP lab of Stanford University in the 1990s, in proposed system using the maxent tagger model into part-of-speech tagger from Stanford parser.POS tagger is more than faster to other available tagger. And maxent tagger model to more faster and accurate to other existing model. It uses the best tokenize method in which each and every word create a token.thats why its increase accuracy of parsing and give the best result of parsing. And easily extract the NER.

MapReduce is a programming model for use expressing distributed computations on huge amounts of data and an execution framework for large-scale data processing on clusters of produce servers.[4] It was originally developed by Google and built on well-known principles in parallel and distributed processing which was already introduce several decades. MapReduce has since enjoyed pervasive adoption via an open-source implementation called Hadoop, whose development was led by Yahoo (now an Apache project).

ApacheTM Hadoop is an open source framework that supports distributed computing. It came into existence from Google's MapReduce and Google File Systems projects. It is a platform that can be used for intense data applications which are

processed in a distributed environment. [5][10] It follows a Map and Reduce programming paradigm where the division of data is the simple step and this split data is fed into the

Manuscript Received Dec, 2014.
***Jigyasa Nigam**, Deptt. of CSE, RGPV University/SRIT Jabalpur, Jabalpur, India,*
***Sandeep Sahu**, , Deptt. of CSE,, RGPV University/SRIT Jabalpur, Jabalpur, India,*

distributed network for processing. The processed data is then integrated as a whole. Hadoop also provides a defined file system for the organization of processed data like the Hadoop Distributed File System(HDFS).[5][10] The Hadoop framework takes into account the node failures and is automatically handled by it. This makes hadoop really flexible and a versatile platform for data intensive applications. The answer to growing volumes of data that demand fast and effective retrieval of information lies in engendering the principles of data mining over a distributed environment such as Hadoop. This not only reduces the time required for completion of the operation but also reduces the individual system requirements for computation of large volumes of data. Distributed Computing is a technique aimed at solving computational problems mainly by sharing the computation over a network of interconnected systems. Each individual system connected on the network is called a node and the collection of many nodes that form a network is called a cluster.

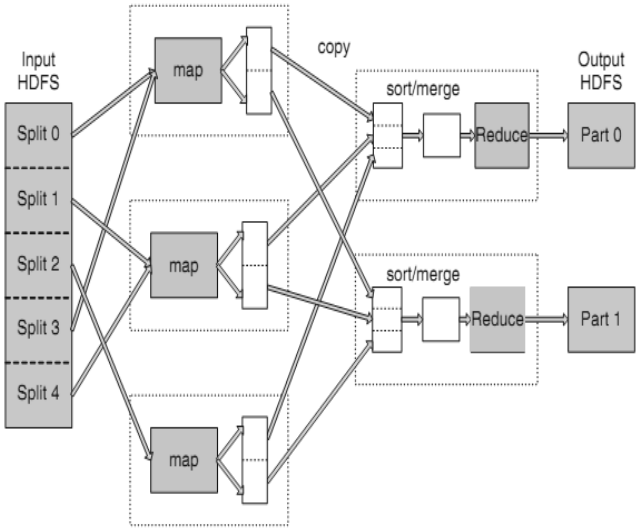


Fig. 2.1 The MapReduce framework

In this fig. shows the input data split equally and apply map function on these files.map function make a (key, value) pair to split data. After that they are combine ,and sort and this sorted pair reduce function are apply .reduce function reduce the size of file with maintain the accuracy and give the final MapReduce result.

III. PROPOSED DISTRIBUTED PARSING SYSTEM

In order to extract information from huge amount of text file. This paper propose to use of distributed environment in which MapReduce programming apply with StanfordPOS-tagger NLP system.[1][2] In figure 3.1 show what system propose in this paper. Figure show how the huge amount of input data to access in hadoop distributed file system with the use of MapReduce programming. Maxent tagger model of Stanford POS-tagger system used by propose system is loaded into hadoop file system because of all mapper function to share it for tokenize the sentences. In this propose system architecture all input file store in master server and master server distributes the files to slave servers. after distribution apply the mapper function and Stanford POS-tagger in each file. Mapper function separate the sentences into key and value form and with the help of maxent tagger model of stanfordPOS-tagger system tokenize the sentences. And in this tokenized sentences recognize name entity form and apply reduce method. In this paper use maxent tagger model because of this model is tokenize the each and every word of sentence that's why we retrieve the more accurate or fast desire result. Maxent tagger model use the best parsing method to all Existing model [8].

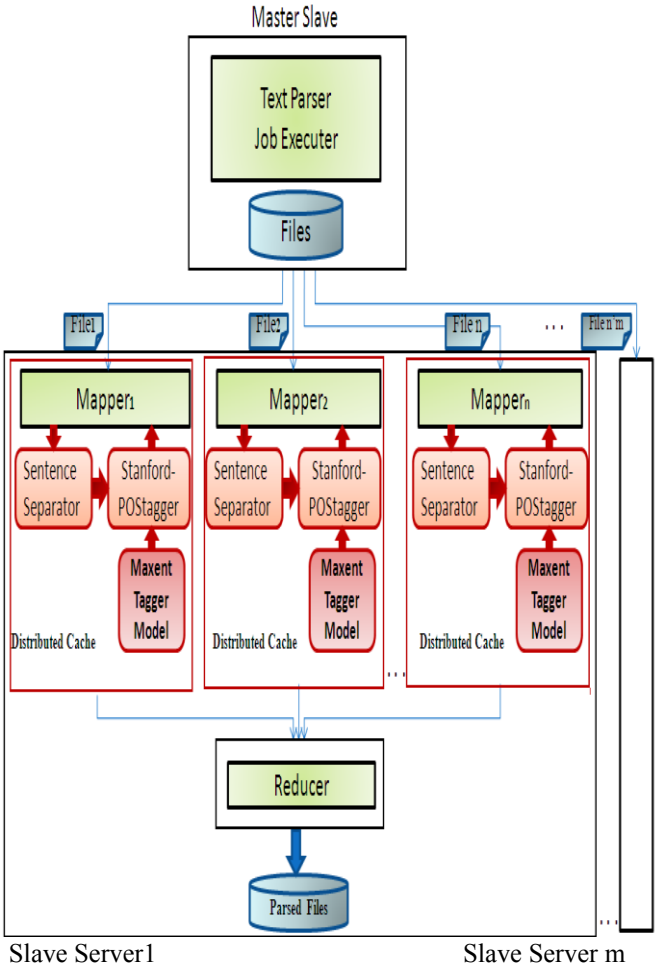
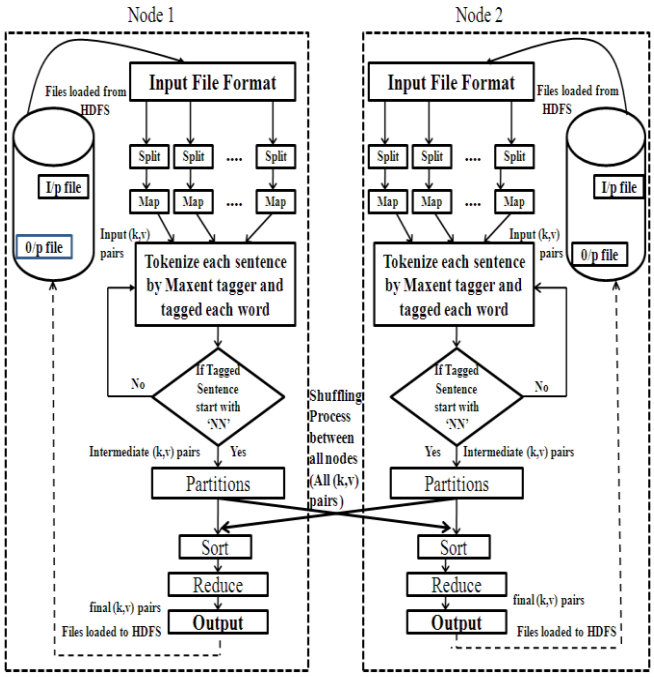


Fig. 3.1 Proposed work architecture

Reducer method reduces the result or length of output file. Output file is also store in master server data storage.



3.2 Flowchart of proposed method

Show proposed system pseudo code in below

Class Mapper

1. StanfordParser.setModel(MaxentTagger model)
2. Method Map(docid a; doc d)
3. sentences = MaxentTagger.tokenizeText(d);
4. sentence1<-sentences.Startwith("NN");
5. Emit(sentence1,d.id)

Class Reducer

1. Method Reduce(d.id ,iterable(sentence1)
2. For each sentence1 s1
3. sum=sum+s1;
4. emit(d.id ,sum)

Proposed system have some advantages, first it reduces the time of parsing then to legacy system because in legacy system tokenize file one by one.

Second, it reduces the size of output file through which we can easily recognize how many times, the particular word uses in all files. Because it provides the count with every word.

Third we can easily modify it with replacing another parsing method.

IV. CONCLUSION

This paper proposes the approach for processing of huge amount of text in distributed environments with MapReduce programming in which Stanford POS-tagger parser applies for name entity recognition. Advantage of propose system, it is less time consuming then to legacy system.

For future work this System evaluate for show the relationship between name entity and in addition optimized technique for parsing in distributed environment.

V. REFERENCES

[1] James J. (Jong Hyuk) Park et al. (eds.), Mobile, Ubiquitous, and Intelligent Computing,Lecture Notes in Electrical Engineering 274,DOI: 10.1007/978-3-642-40675-1_41, © Springer-Verlag Berlin Heidelberg 2014

[2] Kim, J., Lee, S., Jeong, D.-H., Jung, H.: Semantic Data Model and Service for Supporting Intelligent Legislation Establishment. In: The 2nd Joint International Semantic Technology Conference (2012)

[3] Klein, D., Manning, C.D.: Accurate Unlexicalized Parsing. In: Proceedings of the 41st Meeting of the Association for Computational Linguistics, pp. 423–430 (2003)

[4] Dean, J., Ghemawat, S.: MapReduce: simplified data processing on large clusters. In: OSDI, pp. 137–150 (2004)

[5] HDFS (hadoop distributed file system) architecture(2009),http://hadoop.apache.org/common/docs/current/hdfs-design.html

[6] Seo, D., Hwang, M.-N., Shin, S., Choi, S.: Development of Crawler System Gathering Web Document on Science and Technology. In: The 2nd Joint International SemanticTechnology Conference (2012)
Morphological features help POS tagging of unknown words across language varieties

[7] Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the ACL, pages 425–432,Sydney, July 2006. c2006 Association for Computational Linguistics

[8] R. Agrawal and R. Srikant. Fast algorithms for mining association rules. In Proceedings of the 20th International Conference on Very Large Databases (VLDB-94), pages 487–499, Santiago, Chile, Sept. 1994.

[9] en.wikipedia.org/wiki/Information_extraction

[10] Shvachko,K. Yahoo!,Sunnyvale,CA,USA Hairong Kuang ; Radia, S. ; Chansler, R.Mass Storage Systems and Technologies (MSST), 2010 IEEE 26th Symposium on E-ISBN :978-1-4244-7153-9

Jigyasa Nigam is Students of MTech in Department of computer science and engineering from Shri Ram institute of technology (SRIT), Jabalpur, M.P.She has done BE in Department computer science and engineering from Jai Narayan college of technology(JNCT), Bhopal, M.P. in 2007. she has also done Polytechnic Diploma in Computer science from Indira Gandhi womens polytechnic(IGWP), Chhindwara M.P. in 2004..S he has a four years experience in teaching field.Her Areas of interest are Hadoop, Data Structure,Analysis & Design of Algorithm..

Sandeep sahu is Assistant Professor & Head, P. G. Dept. of Computer Science & Engineering, Shri Ram institute of technology (SRIT), Jabalpur, M.P He has done MTech. in computer Science & engineering from IIT Guwahati in 2011.He has done BE in computer science & engineering from (Smrat Ashok technology and institute) SATI, Vidisha,M.P. His Research Areas are: Manet, WSN, Computer Networks.

