International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 09 Issue: 05 | May 2022 www.irjet.net p-ISSN: 2395-0072

Extraction and Retrieval of Web based Content in Web Engineering

K. Ananthi[1], Dr. Nancy Jasmine Goldena[2]

[1]PG Scholar, Department of Computer Applications and Research Centre, Sarah Tucker College (Autonomous), Tirunelveli, Tamil Nadu, India
[2]Associate Professor, Department of Computer Applications and Research Centre, Sarah Tucker College (Autonomous), Tirunelveli, Tamil Nadu, India

---

Abstract - *The rapid and wide-ranging dissemination of data and information through the internet has resulted in a large dispersion of normal language textual holdings. In the current environment for selecting, distributing, and retrieving a vast supply of information, excessive attention has emerged. Processing large amounts of data in a reasonable amount of time is a significant difficulty and a critical need in a variety of commercial and exploratory industries. In recent years, computer clusters, distributed systems, and parallel computing paradigms have been more popular, resulting in significant advances in computing presentation in data-intensive applications like as Big Data mining and analysis. NLP is one of the key aspects that may be used for text explanation and initial feature extraction from request areas with high computing resources; as a result, these duties can outperform comparable designs. This research presents a distinct architecture for parallelizing NLP operations and crawling online content. The mechanism was discovered using the Apache Hadoop environment and the MapReduce programming paradigm. In a multi-node Hadoop cluster, authentication is done utilizing the explanation for extracting keywords and crucial phrases from online articles. The proposed work's findings indicate greater storage capacity, faster data processing, shorter user searching times, and correct information from a huge dataset stored in HBase.*

Key Words: Natural Language Processing, Hadoop, Text Parsing, Web Crawling, Big Data Mining, HBase.

1. INTRODUCTION

Big Data is described as a collection of datasets with a scale that makes it difficult to govern and distribute information using standard technologies (i.e., management of N-dimensional data circles by necessary text files and SQL files). These challenges create Big Data monoliths [1] as plentiful as secondary information systems (Pollock, 2013a), resulting in primary failure for the largest isolated and public data providers in which a smaller volume of data does not imply easier administration (Akers, 2013) [2]. Information may be extracted using a variety of approaches utilizing NLP. Text mining (Pande et al., 2016) [3] is one of the most effective ways for extracting information from text data. The accuracy, slot error rate, F-measure, and recall of information extraction are all elements to consider. The technology of Information Retrieval (IR) and Information Management (IM) arose from the emergence of efficient information. [4] Extraction (IE) 2018 (Sonit Singh). In IE systems, natural language is used as an input, and it provides structured information according on certain criteria that may be used to a specific application. Data mining techniques handle large datasets to extract key patterns from data; social networking sites provide a large number of datasets for its practice, making them ideal candidates for mining data using data mining tools (Charu Virmani et al., 2017) [5]. As a result, web mining or data mining provides critical insight to a social network in order to correctly develop and interact in a user-friendly manner. Web data mining based on Natural Language Processing (Yue Chen, 2010) [6] involves knowledge representation. The corpus test is made up of 400 words retrieved from the Web new corpus, and this sort of information includes descriptions of events, relationships, and object properties. At the semantic level, this model exemplifies the web page, and its knowledge structure is scalable.

NLP is a theoretically grounded set of computer approaches for analysing and demonstrating human language. It allows computers to do a wide range of common language-related activities at all levels, ranging from analysing and part-of-speech (POS) categorization to machine conversion and conversation schemes. In addition to methods, deep learning designs comprise previously accomplished imposing indicators of success in realms like processor vision and design thankfulness. Following this trend, contemporary NLP research is increasingly focusing on the use of novel deep learning algorithms (Mikolov et al., 2010; Mikolov et al., 2013) [7, 8] and deep learning methodologies (Socher et al., 2013) [9]. Multilevel programmed feature representation learning is possible with deep learning. Traditional machine learning-based NLP systems, on the other hand, rely heavily on human-crafted structures. Handcrafted arrangements take time to create and are usually flawed.

2. RELATED WORK

Barbosa et al., (2015) discussed several strategies for information extraction [10]. And he went on to discuss knowledge inferencing, which entails using a variety of data sources and extraction techniques to check existing and

current knowledge. Deep natural language processing utilizing machine learning approaches, large-scale probabilistic reasoning, data cleansing using integrity techniques, and using human experience for domain knowledge extraction are the four inferencing methodologies.

As a result, such algorithms may be employed in a variety of systems to extract information from online phrases. Chandurkar et al., (2017) presented a Question Answering (QA) system that integrates the domains of information retrieval (IR) and natural language processing (NLP) [11]. The QA system was based on the UI's intended topic and inquiry (User Interface). To encode the XML document, the TREC 2004 dataset is used. By adding a text box to the system, the user may ask his or her own inquiry. The exact DBpedia page is based on the target subject, as well as the Stanford Dependency Parser, which incorporates the other process for extracting the target or focused topic in the inquiry. However, it is entirely dependent on the factoid question classifier in Python.

Florence et al., (2015) introduced the summarizer system, which is based on online document semantic analysis [12]. The system constructs the summary using Cluster values in Resource Description Framework (RDF) space. Based on the original length of the papers, this system may generate short and lengthy summaries. Finally, the clustering method extracts the subject, object, and verb (SOV), which may then be grouped together using the RDF triples produced. The chosen important phrases are then extracted using the Sentence Selection (SS) algorithm for the final summary. In several NLP tasks such as Semantic Role Labeling (SRL), Named-Entity Recognition (NER), and POS tagging, Collobert et al. (2011) discovered that a simple in-depth learning approach outperforms most complex techniques. Meanwhile, several difficult deep learning-based processes to address tight NLP obligations have been proposed [13]. Goldberg (2016) easily exposed the key ideology centered on smearing neural networks to NLP in a pedagogical way in his research. The researchers hoped that their effort will provide readers a more comprehensive picture of current research in this field [14]. NLP techniques, according to Azcarraga et al., (2012), are the foundation of language-based solutions that often deliver more precise outcomes using statistical approaches; nevertheless, they are computationally extremely exclusive [15].

Pavithra et al., (2013) investigated the wrapper induction technique for data extraction. The system is advanced by the usage of a combination of XSLT and DOM through XML. For web-based applications, XML methods are quite useful. According to the content, this approach provided efficient information extraction and chained wrapper modification. It enhanced usability and flexibility [16]. Sridevi et al., (2016) proposed a Viterbi technique for constructing medical corpus data, which included information extraction from clinical language based on context, allowing clinicians to make quicker and improved conclusions. It also improved the treatment's quality [17]. Wu Wei et al. (2013) developed a new extraction rule language that described the integrated logic for data integration, information extraction direction-finding, and web page extraction. A source data object is used to represent and encapsulate a web data area, including information records and objects. The system enables users to construct sophisticated target data entities using XML, explains the structure of the target data entity, and previously approved integration scripts to transform and map information taken from source to target data objects [18]. Jindal et al. (2013) developed a similar NLP system based on the Learning Based Java (LBJ) methodology [19]. In addition to using Charm++, Rizzolo and Roth (2010) used a similar software design approach [20]. Exner and Nugues (2014) described the Koshik multi-language NLP platform, which was designed for large-scale processing, such as the analysis of unstructured natural linguistic materials spread throughout a Hadoop cluster [21]. Text tokenization, dependency parsers, and co-reference solvers are among the algorithms it offers. Using a Hadoop distributed architecture and its Map Reduce program design methodology to professionally and easily grow by adding low-cost commodity hardware to the cluster is a competency advantage. Barrio and Gravano (2017) explained how to extract information schemes in natural language script to identify classified information. When opposed to a likely finished native linguistic document, knowing controlled technique allows for much more affluent enquiring and data mining. However, outside illuminating a competency of an extraction technique with text groups of fragile attention, extracting a substance is a computationally pricey mission.

The study focuses on a well-known family of text groups, the so-called deep-web text groups, whose insides aren't crawlable and can only be accessed by querying. For efficient content extraction from deep-web text groups, there is a crucial step [22]. Wang and Stewart (2015) investigated geographic information science by modeling spatial dynamics discovered in spatiotemporal content collected from the Internet, especially unconstructed facts such as online news headlines. Taking into account both spatial and temporal data from a collection of Web forms enables us to create a rich exemplification of geographic elements identified in the text, such as where, when, and what occurred. This study looks at how part ontologies operate as a key component in a semantic material abstraction approach. They demonstrated how ontologies might be used in teaching grammatical process material around hazard spatially and proceedings complement abstraction with semantics by combining them with traditional linguistic gazetteers [23].

Match and Avdan (2018) presented a method for removing obstacles (such as typographical errors, spelling mistakes, and inappropriate arrangements) that impede the accuracy of a geocoding approach. The data of an address is examined using NLP methods in the present methodology. Both the Match Rating Compute Method and the Levenshtein Distance

Procedure are unaffected by misspellings, abbreviations, or omissions. As a result, the addresses are rearranged into a precise order. Calculate the technique's properties based on the results of the geocoding operation. A test dataset comprising Eskishehir elementary school addresses is created. Both before and after the calibration method, the geocoding technique is validated using a current sample of addresses [24]. Glauber and Claro (2018) described strategies to guarantee that all relevant primary research is included. Even if we strew our inquiry threads over five files, an EMNLP conversation distils the most important principle.

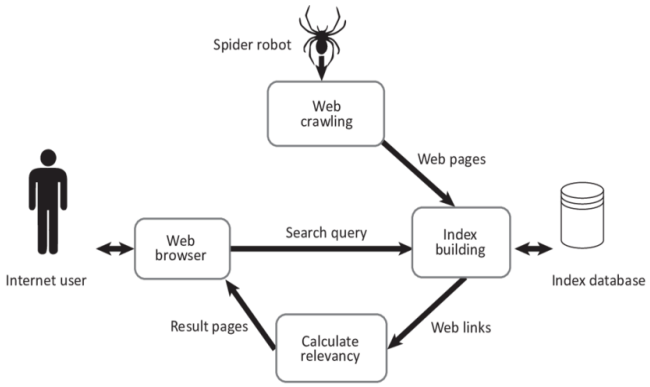

Fig -1: Architecture diagram

3. PROPOSED METHODOLOGY

Users may seek for trends in interactive data mining from a variety of perspectives. Because it is difficult to know what might be disclosed inside a database, storing large amounts of data, and processing data quickly, the data mining approach must be interactive. Different types of data are stored in databases and data warehouses. It is unthinkable that one system could harvest all of these different forms of data. As a result, the different data mining systems must be interpreted for varied data types.

3.1 Description

The framework accepts the input name as a URL and extracts all forms of material from web sites, which is then shown in the textbox. The phrases from noun or verb clauses are extracted using Parts-of-Speech (POS). It assigns data to each token that has been identified. Map and Reduce are two critical jobs in the Map-Reduce system. The individual components are smashed into a tuple in this case. In this case, the map translates a particular collection of data into a variable set.

The activity of the guide or mapper is to put the knowledge into practice. The information is stored as a document or registry in the Hadoop record framework (HDFS) for the most part. The Reducer's job is to process the data that comes in from the Mapper. Following the preparation, it gives another yield technique, which will be used in the HDFS.

3.2 Modules

- Dataset Collection
- Map Reducer
- Mapper tasks
- Reducer tasks
- Hadoop Storage

3.3 Module Description

a. Dataset Gathering

The Input name is supplied as a URL in this Module, and then various contents are taken from web sites and shown in the textbox. Web Harvesting (WH) and Web Scrawling are two related methods used by websites to direct user queries to their site. The URL is used to collect the processed and extracted dataset (uniform resource allocation).

b. Pre-Processing

Tokenization is the process of breaking down lengthy strings of material into smaller tokens. Larger chunks of material may be tokenized into sentences, which can then be tokenized into words, and so on. Additionally, much prepping is done after a piece of writing has been properly tokenized.

Tokenization is referred to as a lexical study or a content division. Tokenization refers to the breakdown of a large chunk of text into smaller bits (e.g., passages or sentences), while division refers to the breakdown technique that just produces words.

c. WH-NLP-POS Tagging Algorithm

A web harvesting software finds websites with specific material targeted at a certain web harvest request. The web data is collected by the web harvesting application, which also includes a link that takes the user to the company's website. This information is then indexed by well-known search engines like Yahoo and Google, making it easier to find it in future searches. Increasing visibility via web harvesting is critical for a company's online expansion and growth. The use of web harvesting is a major aspect in generating online company income. As a result, there will be a higher chance that a prospective consumer will find your website via an internet search.

As a result, obtaining outside support in their attempts to expand their online presence has become critical for corporate success. Finally, organizations should look for the most cost-effective and user-friendly web harvesting software available to guarantee that their efforts are always rewarded with favourable results. The terms "web scraping" and "web harvesting" are interchangeable, while "web harvesting" refers to crawling several sites and extracting a

specific collection of data. It's also known as targeted web crawling.

FMiner is an important tool for web harvesting. Multiple site URLs may be used as the beginning URL in an extraction project file, and the extracted results can be stored to the same database. The application may then collect page contents from numerous sites constantly and gradually using the "schedule" and "incremental" capabilities.

d. Natural Language Processing (NLP)

Natural language processing for big data may be used to automatically locate important information and/or summarize the content of documents in vast quantities of data for collective understanding. Natural Language Processing (NLP) is a technique for analysing understandable text produced by humans for the purposes of language processing, artificial intelligence, and translation.

There are several NLP approaches for dealing with issues such as text corpus collection and storage, as well as text analysis, in order to effectively and precisely analyse the text. NLP techniques benefit and acquire experience from linguistics and artificial intelligence research. Intelligence, machine learning, computational statics and other sciences.

However, at recent times, because of the explosion of information, the utilization of traditional NLP faces many challenges such as the volume of unstructured and structured data, accuracy of the results and velocity of processing data. In addition, there are so many slangs and indefinite expressions used on social media networking sites, which give NLP pressure to examine the meanings, which may also be difficult for some people.

Furthermore, nowadays, people heavily depend on search engines such as Google and Bing (which use NLP as their core technique) in their daily study, work, and entertainment. All of the above-mentioned factors encourage computer scientists and researchers to find more robust, efficient and standardized solutions for NLP.

e. NLP Tokenization

Tokenization may be described as the process of splitting the text into smaller parts called tokens, and this is deliberated as a crucial step in NLP. The process of slicing the given sentence into smaller parts (tokens) is called as tokenization. In general, the given raw text is tokenized based on a group of delimiters. Tokenization is used in tasks such as processing searches, spell-checking, identifying parts of speech, document classification of documents, sentence detection etc.

- **Simple Tokenizer** – Tokenizes the available raw text using character classes.
- **Whitespace Tokenizer** – Uses whitespaces in order to tokenize the given text.
- **Tokenizer ME** – Converts raw text into separate tokens. It uses Maximum Entropy to make its decisions.

Steps

- Tokenize the sentences
- Print the tokens
- Instantiating the respective class.

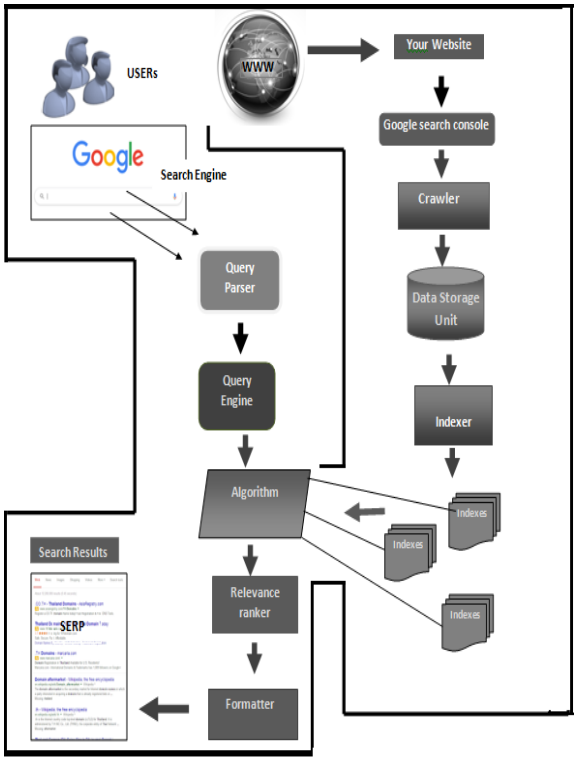

Fig -2: Data flow diagram

f. POS Tagging

The Parts of Speech of a given sentence can be detected using OpenNLP and then can be printed. Instead of full name of the parts of speech, OpenNLP uses short forms of each part of speech. Then, a map in addition to reduce phases run in slots implies that each node could run beyond one Map or else Reduce task in matching; generally, the slots quantity is correlated through cores quantity present in a specific node.

g. Reducer tasks

Reducer is a phase in Hadoop which comes after Mapper stage. The yield of the mapper is provided as the contribution for Reducer which proceeds and creates another arrangement of creation, which will be put away in the HDFS. Reducer first handles the in-between values for individual key produced as a result of the map function in addition to further providing an output. Where - output information of a map phase, - number of map tasks, -single reduce task.

4. RESULTS AND DISCUSSION

The expected system scalability performance is provided in a manner similar to the earlier authentication. On the other hand, the study expanded the dataset from 10,000 to 20,000 web pages and documents and evaluated the time of dispensation for Keyword Extractor Module (KEM) implementation on the text content of a previously investigated dataset using a Nutch-based crawler. The Web Extraction technique is shown in Figure 2. As a result, no additional external network access will be required for keyword extraction or keyword extraction. This strategy eliminates bottlenecks while also removing the need for process parallelization. Figure 3.

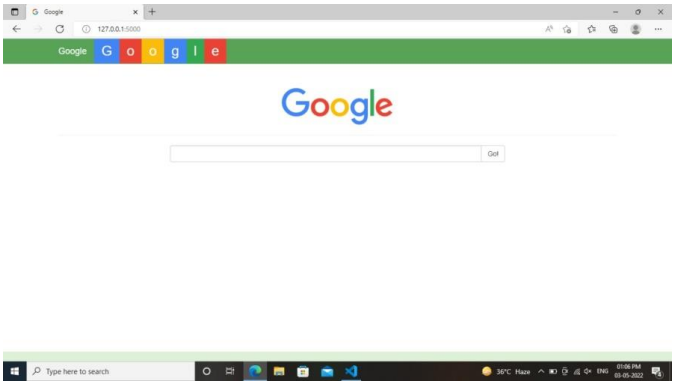

Fig -3: Search Engine

Method of processing with respect to the version anticipated in Nesi et al., (2015), several advances have defined the possible solutions for these overall changes to a code as well as a test formation, both in terms of time presentations and scalability. The Hadoop cluster design used in the testing was assessed in a variety of configurations, ranging from two to five nodes. By utilizing Hadoop HDFS, each node is a Linux 8-core workstation. Hadoop allows completing a rebalancing of deposited blocks across the cluster active nodes to avoid data integrity problems and disappointments due to decommissioning and recommissioning of cluster nodes [26].

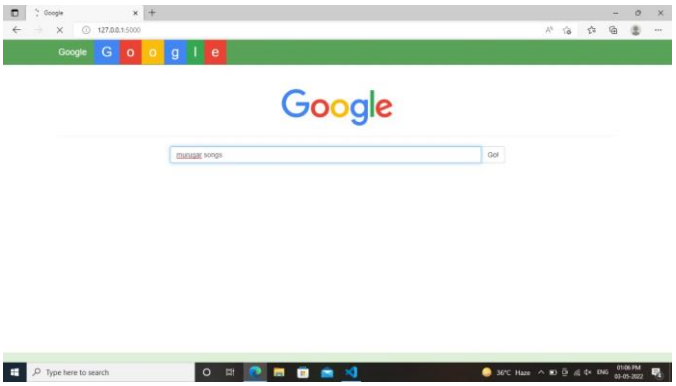

Fig -4: Content Search

After providing a URL, all content categories are extracted from web sites and shown in a textbox. NLP is used to pre-process and organize the data collected. The document's keywords are extracted. The keyword and the keyword exchange information. The document's words are recognized. The report is determined using the neural network between the keyword and the words inside. Only if the model is detected is the text chosen. Figure 4 depicts the Content Extraction Testing.

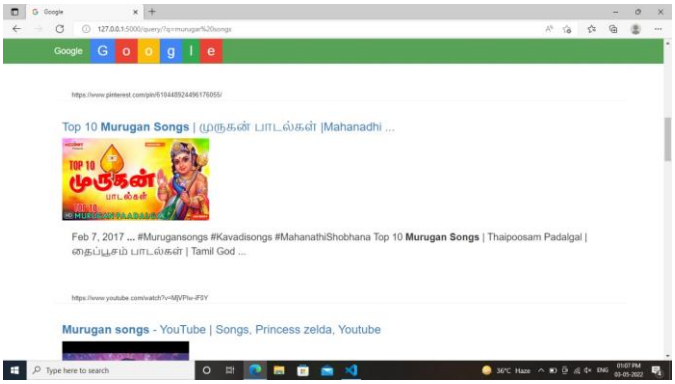

Fig -5: The Testing of Content Extraction

The MapReduce concept allows for speculative job execution and is designed to provide redundancy for fault tolerance. As a result, the Job Tracker may be required to postpone fizzled or otherwise terminated tasks, which may affect the time it takes to complete the entire operation. As a result, the ideal prepared conditions have been picked for execution correlation among a few test occasions that have been lead for each hub configuration. The amount of work required to retry failed or eliminated jobs must be lowered using the existing system. Approximately 3.5 million documents were retrieved from the entire examination dataset.

Fig -6: The Map-Reducer Process

As an examination term, running the on a single non-Hadoop workstation, the same CiteSeerExtractor, a RESTful API application on a comparable corpus dataset took roughly 115 hours. The Map-Reducer technique is depicted in Figure 5. The fact that the standalone CiteSeerExtractor, a RESTful API application, is not optimized for multi-threading may be a possible explanation for this important presentation opening. Despite the way that Hadoop's Map-Reduce adaptation may take benefit of MapReduce configuration choices that describe the largest number of guides and

reduce errand spaces, all of this can be accomplished by using multi-center innovation even on a single hub cluster. The Hadoop Storage data set is shown in Figure 6.

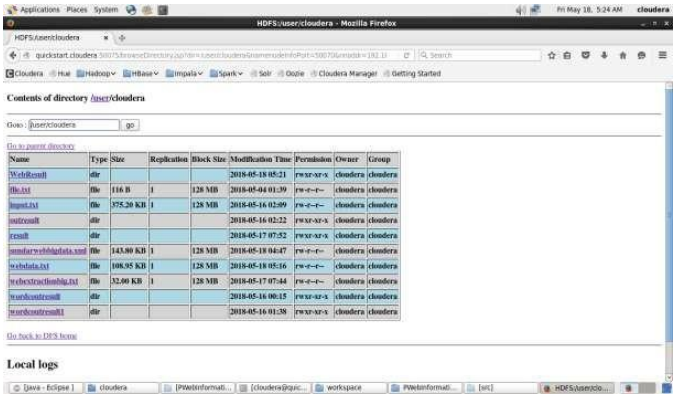

Fig -7: The Hadoop Storage

The MapReduce contains two essential tasks, namely Map and Reduce. The map consists of a group of data and changes it into an alternative group of data, where separate essentials are fragmented down into the tuple. The map or mapper's work is to develop the input data. Generally, the contribution data is present in the procedure of directory or file and stored in the Hadoop file system (HDFS). Hadoop assigns Map and Reduce tasks to the appropriate machines in the cluster. The diagram accomplishes all data-passing details, such as assigning duties, checking job completion, and moving data from the group to the nodes. The majority of computation occurs on nodes on original recordings, which reduces network circulation. After completing the assigned tasks, the team gathers and reduces the data to produce a sufficient result before returning it to the Hadoop server.

5.1 Web Information Extraction validation results

a. Precision

Our proposed WPE_NLP methods have the highest precision compare than Existing UPE_DC method.

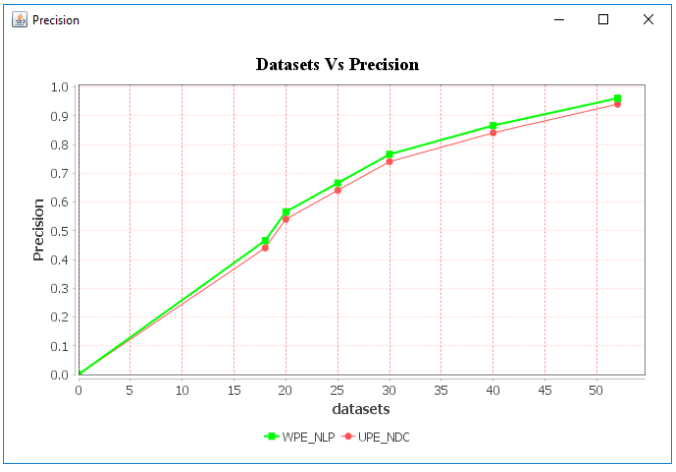

Fig -8: Datasets compare with precision

b. Recall

Our proposed WPE_NLP methods have the highest Recall compare than Existing UPE_DC method

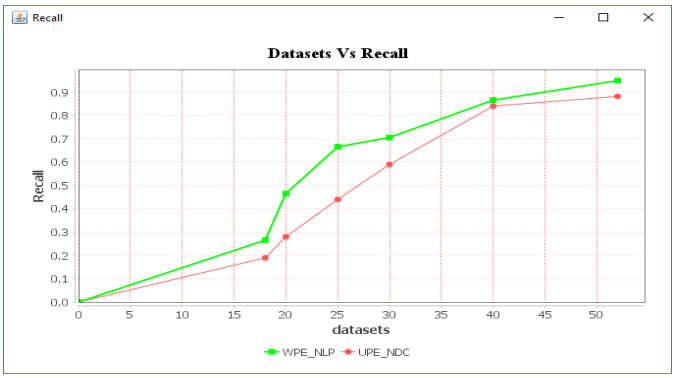

Fig -9: Datasets compare with recall

c. F-Measure

Our proposed WPE_NLP methods have the highest F-Measure compare than Existing UPE_DC method

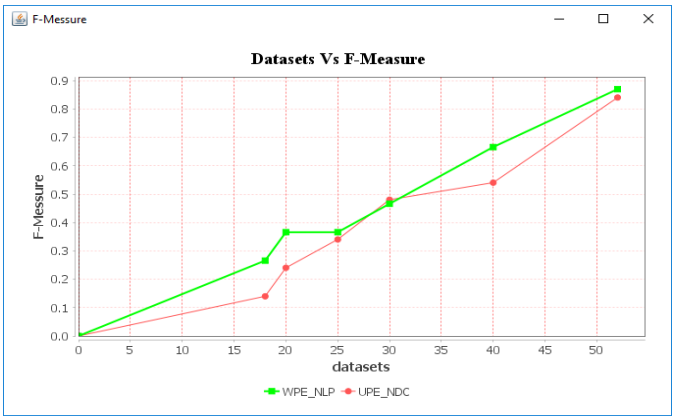

Fig -10: Datasets compare with F-Measure

3. CONCLUSION AND FUTURE WORK

The findings of this investigation point to web-based datasets and the benefits of Big Data. It is critical to develop a method that ensures the use, administration, and re-use of information sources, including online URL information, as well as the collection of useful URL and services across the country. It's critical to decide the strategy to use for Information Extraction using Neural Natural Language Processing (IE-NLP). Hadoop with MapReduce may be used to improve analytic processing. The basics of MapReduce software are developed by the open-source Hadoop backdrop, according to the present study. Hadoop's excellent context raises the management of large amounts of data over the distribution process and reacts quickly. Big data processing in the future will include multiple types of data, such as unstructured, semi-structured, and structured data.

REFERENCES

[1] Pollock, R., 2013a. Forget Big Data; Small Data is the Real Revolution d Open Knowledge Foundation Blog. http://blog.okfn.org/2013/04/22/forget-big-datasmall-data-is-the-real-revolution.

[2] Akers, K.G., Feb. 2013. Looking out for the little guy: small data curation. Bull. Am. Soc. Inf. Sci. Technol. 39 (3), 58-59.

[3] Pande, V., & Khandelwal. Information Extraction Technique: A Review. IOSR Journal of Computer Engineering (2016) 16-20.

[4] Sonit Singh, Natural Language Processing for Information Extraction, IEEE publications (2018) 1-24.

[5] Virmani, C., Pillai, A., & Juneja, D. Extracting Information from Social Network using NLP. International Journal of Computational Intelligence Research (13) (4) (2017) 621-630.

[6] Chen, Y. Natural Language Processing in Web data mining. IEEE 2nd Symposium on Web Society, 2010, 388-391.

[7] Mikolov, T., I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," in Advances in neural information processing systems, 2013, pp. 3111–3119.

[8] 8. Mikolov, T., M. Karafiat, L. Burget, J. Cernock´y, and S. Khudanpur, "Recurrent neural network based language model." in Interspeech, vol. 2, 2010, p. 3.

[9] Socher, R., A. Perelygin, J. Y. Wu, J. Chuang, C. D. Manning, A. Y. Ng, C. Potts et al., "Recursive deep models for semantic compositionality over a sentiment treebank," in Proceedings of the conference on empirical methods in natural language processing (EMNLP), vol. 1631, 2013, p. 1642.

[10] Barbosa, D., Wang, H., & Yu, C. (2015). Inferencing in information extraction: Techniques and applications. 2015 IEEE 31st International Conference on Data Engineering, 1534- 1537.

[11] Chandurkar, A., & Bansal, A. (2017). Information Retrieval from a Structured Knowledgebase. 2017 IEEE 11th International Conference on Semantic Computing (ICSC), 407- 412.

[12] Florence, A., & Padmadas, V. (2015). A summarizer system based on a semantic analysis of web documents. 2015 International Conference on Technologies for Sustainable Development (ICTSD), 1-6.

[13] Collobert, R., J. Weston, L. Bottou, M. Karlen, K. Kavukcuoglu, and P. Kuksa, "Natural language processing (almost) from scratch," Journal of Machine Learning Research, vol. 12, no. Aug, pp. 2493–2537, 2011.

[14] Goldberg, Y. "A primer on neural network models for natural language processing," Journal of Artificial Intelligence Research, vol. 57, pp. 345–420, 2016.

[15] Azcarraga, A., M. David Liu, and R. Setiono, "Keyword Extraction Using Backpropagation Neural Networks and Rule Extraction," in Proc. of IEEE World Congress on Computational Intelligence (WCCI), Brisbane, Australia, June 2012.

[16] Pavithra, Monisa, & Ramya. (2013). A Design of Information Extraction System. International Journal of Advanced Research in Computer Science, 4 (8), 109-111.

[17] Sridevi, & Arunkumar. (2016). Information Extraction from Clinical Text using NLP and Machine Learning: Issues and Opportunities. International Journal of Computer Applications, 11-16.

[18] Wei, W., Shi, S., Liu, Y., Wang, H., Yuan, C., & Huang, Y. Extraction Rule Language for Web Information Extraction and Integration. 2013 10th Web Information System and Application Conference (2013), 65-70. doi:10.1109/wisa.2013.21

[19] Jindal, P., D. Roth, and L.V Kale, "Efficient Development of Parallel NLP Applications," Tech. Report of IDEALS (Illinois Digital Environment for Access to Learning and Scholarship), 2013.

[20] Rizzolo, N., and D. Roth, "Learning-Based Java for Rapid Development of NLP Systems." In Proc. of the International Conference on Language Resources and Evaluation (LREC), 2010.

[21] Exner, P. and Nugues, P., "KOSHIK - A Large-scale Distributed Computing Framework for NLP," in Proc. of the International Conference on Pattern Recognition Applications and Methods (ICPRAM 2014), pp. 463- 470, 2014.

[22] Pablo Barrio and Luis Gravano. Sampling strategies for information extraction over the deep web. Information Processing and Management 53 (2017) 309–331.

[23] Wei Wang and Kathleen Stewart. Spatiotemporal and semantic information extraction from Web news reports about natural hazards. Computers, Environment and Urban Systems 50 (2015) 30–40

[24] Dilek Küçük Match and Uğur Avdan. Address standardization using the natural language process for improving geocoding results. Computers, Environment and Urban Systems, 2018.

[25] Rafael Glauber, Daniela Barreiro Claro, A Systematic Mapping Study on Open Information Extraction, Expert Systems with Applications (2018), doi: 10.1016/j.eswa.2018.06.046

[26] Nesi, P., G. Pantaleo, and G. Sanesi, "A Distributed Framework for NLP-Based Keyword and Keyphrase Extraction from Web Pages and Documents," in Proc. of 21st Int. Conf. on Distributed Multimedia Systems