Proceedings of the Second International Conference on Research in Intelligent and Computing in Engineering pp. 135–141 DOI: 10.15439/2017R86
ACSIS, Vol. 10 ISSN 2300-5963

Metadata based Text Mining for Generation of Side Information

Shraddha S. Bhanuse[1], Shailesh D. Kamble[2], NileshsinghV. Thakur[3], Akshay S. Patharkar[4]
[1]Software Engineer, Mobisoft Infotech, India
[2]Computer Science & Engineering, Yeshwantrao Chavan College of Engineering, India
[3]Computer Science & Engineering, Prof Ram Meghe College of Engineering & Management, India
[4]Computer Technology, K.D.K. College of Engineering, India
[1]shradha.bhanuse@gmail.com, [2]shailesh_2kin@rediffmail.com,
[3]thakurnisvis@rediffmail.com, [4]akshay.patharkar7@gmail.com

***Abstract*—Text mining is knowledge analyzing technique to find a pattern. The side information is also called as metadata in most of the metadata based text mining applications. The side information consisting of large data in terms of weblogs, metadata, and non-textual data i.e. image/video, etc. This large data present in the unprocessed form which cannot be used for further text mining. Therefore, metadata based text mining algorithms are used to mine the useful information. In this paper, the proposed approach uses the different kind of pre-processing steps i.e. splitting, tokenize, steaming, parsing and chunking. For generating the side information i.e. title, name, affiliation, email address, place etc. a natural language processing (NLP) is used. To achieve the effective clustering, the proposed approach uses a classical partitioning method with a probabilistic model. The proposed approach is compared in terms of time required for mining of words, accuracy, and efficiency. The presented result shows that, the proposed approach performs better in terms of accuracy and running time. In future, a Security is provided for metadata based side information generation using Intrusion Detection System (IDS).**

***Index Terms*—Text Mining; Metadata; Text Mining; Side Information; Natural Language Processing; Classical Partitioning; Clustering**

I. Introduction

The metadata based text mining means to retrieve the useful information from the large dataset. How to analyze the knowledge from unstructured texts [1-2] is the main research component present in text mining called as text data mining (TDM) [3-4]. The process of text mining is similar to data mining process; difference is that the data mining tools are used to handle structured data whereas text mining tools are used to handle unstructured data sets ex. HTML files or any full-text documents etc. [5]. Text Mining is useful for creating new or unknown information from different available resources i.e. from different databases available on www/internet media. Text Mining is the research area of computer science which consists tough links with NLP, data mining, machine learning, information retrieval and knowledge. To find and examine an interesting mining request to extract use full information from shapeless textual data through the patterns. The Preprocessing in Text Mining is as shown in Figure 1.

A. Side Information

The text clustering occurs indifferent domains such as the internet media/www, social networking site, etc. [6-8]. A more research work has already been presented on the problem of clustering in text data and retrieve the same information. Still there is a scope in improvisation on problem of clustering text data. The examples of side-information generation are as follows:

(a) Text Document Contains Links: Text document contains some helpful information for mining [9]. It also provides relationships among measurement and others are deliberate. By using specifications that anticipate your paper as one part of the entire proceedings, and not as an independent document.

(b) Meta-data: Web documents contains data of data [10] consists of different kinds of information about the document. For example, ownership, locality, or any temporary information which is useful for mining techniques. In some user or knowledge sharing information contains user tag, which is also very important. This is quite informative.

B. Natural language Processing

NLP is a recent research domain in a computer science and engineering. NLP discovers and analyze that, how the computer system can be used to understand and manipulate natural language text. The aim of researchers is how to collect the information, understand the information by user and use language in specific domain. Therefore, the methodologies are used to develop computer systems to understand and manipulate natural languages for text mining [11-14]. NLP provide large scale disciplines and responsibilities for achieving and expanding the capabilities of text mining, or the extraction of knowledge from shapeless, example is machine-learning paradigm of language processing. NLP [3, 15] algorithms should meet some success in structured or unstructured fields examples are medicine and biochemistry. The various methods of NLP are [12]: Part-of-speech tagging (POS); Tokenization; Splitting; Parsing, and Text chunking aims at grouping adjacent words in a Sentence. The basics of NLP consists of several categories like computer and information sciences, mathematics, linguistics,

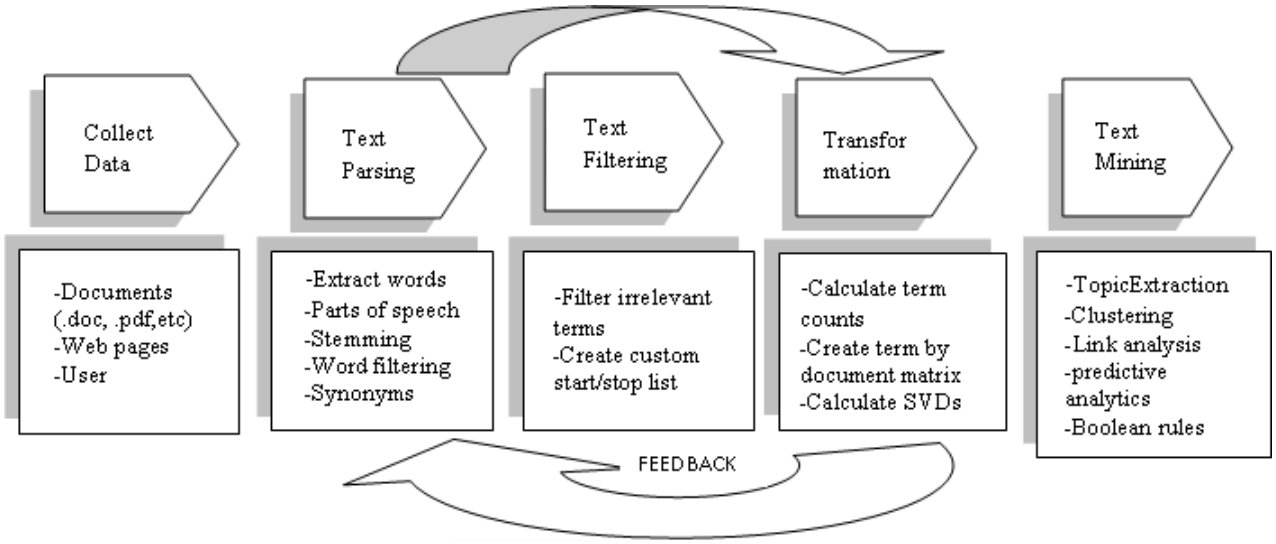


Fig.1. Preprocessing in Text Mining

and electronic engineering, fuzzy logic, robotics and artificial intelligence, psychology, etc. Applications of NLP's are language text processing and summarization, machine intelligence, user interfaces, and cross-language information retrieval (CLIR), artificial intelligence systems, speech recognition, and so on [16-17].

The remainder of this paper is organized as follows: Section II describes the related work on metadata based text mining for side information generation. Section III elaborates the proposed approach for side information generation. Section IV summarizes the experimental results of a proposed approach. Finally, this paper ends with the conclusion and future scope in Section V followed by the references section.

II. RELATED WORK

In metadata based text mining, huge web online collection is the main reason to develop a mechanism to create effective and scalable clustering algorithms used for generating side information [1-3]. The current proposed approaches focus on data processing to maximize the clustering advantage to generate side information. Jain and Dubes [18] proposed an approach for clustering text data with side information. It provides an idea to perform mining process a way to perform the mining process as to maximize the benefits of side information. It uses an algorithm which is a combination of traditional partitioning algorithms with the probabilistic models [1]. The stemming is the process in mining, to reduce different grammatical or word forms of a word like its noun, adjective, verb, adverb etc. Stemming is used for reducing inflectional forms. This paper discusses different methods of stemming and their comparisons in terms of usage, advantages as well as limitations. The basic difference between stemming and lemmatization is also discussed. Yang et al. [9] discussed in the methodologies for developing computing applications that will be flexible and adaptable for users. In this context, however, information retrieval (IR) system used to find location and delivery documents to satisfy user's need. The stemmer's effectively used such as spelling checker, and may vary language to languages. The working of typical simple stemmer algorithm is removing suffixes using a list of frequent suffixes is discussed in [12].One more complex thing is to use morphological knowledge to derive a stem from the words. The proposed approach gives a detailed view of common stemming techniques.

There are many problems in clustering: first, for large clustering, process is too slow and second, that retrieval information as per user's requirement is not improved by classification and clustering techniques. Basically clustering is used to improve predictive search and analysis. The initial browsing technique is nothing but the Document clustering [19-20]. Always there is strong association rule between clustering and its technique. Metadata based text mining is used for feature compression and extraction of reducing dimensionality. Clustering is main challenge when data is in heterogeneous form. We have different types of algorithm and techniques for classification and clustering. In text mining, data pre-processing plays very important role. Aggarwal et al. [8] presented a survey on text data classification and clustering algorithm. For classification and clustering, data is extracted or used from metadata for generating side information. Guha et al. [4] suggested the unknown discovery pattern or identifying interesting pattern in terms of data clustering is used in data mining.

The clustering algorithm CURE which is more accurate to outliers, and identifies clusters [21]. Because cluster having non-spherical shapes. They are wide variances in size and shape. Zhong et al. [22] proposed an effective pattern discovery technique which includes the processes of pattern deploying and pattern evolving. To enhance the effectiveness of modifying the discovered patterns to find an appropriate unknown pattern, Franz et al.[23] proposed an unsupervised and supervised learning, will help to improve the quality of the

clustering effects of both the text metadata and side information.The proposed approach shows the extension of the clustering approach to the metadata based text classification using the side information or generating side information of the text documents. Jivani [24] discussed the purpose of stemming is to reduce different grammatical forms or word forms of a word like its noun, adjective, verb, adverb etc. The goal of stemming is to minimize inflectional forms and sometimes derivationally related forms of a word to a common base form. This paper discusses different methods of stemming and their comparisons in terms of usage, advantages as well as limitations. The basic difference between stemming and lemmatization is also discussed.

III. Proposed Work

The proposed work consists of various steps like preprocessing, extraction,and stemming, etc. these are depicted in the work flow of the proposed work in Figure 2.

- Collection of datasets for data mining- Selecting an appropriate dataset for comparing results and analysis. After selecting dataset some operations can be carried out Example- crawling, filtering, etc.
- Preprocessing can be performed with natural language processing-In which we would be applying natural language processing techniques like splitting the document, tokenizing, part of speech tagging and chunking. The aim of natural language processing (NLP) is to convert human language into a machine understanding form, is easy for computers manipulate data and its meaning.
- The general goal of NLP is to achieve a better and understanding of natural language by use of computers. It is very simple and fast technique for data mining.
- For demonstration of text mining approach, proposed approach develops a mining algorithm with Natural Language for processing-preprocessing of text. Use of Word Net-It is a collection of words in a dictionary form language where dictionary of language containing combination of mining with natural language processing-meanings, senses, etc of words. In this step we will be showing the meaning of searched data.

In Figure 2 all the above work would be important for pre-processing. Text mining uses NLP for demonstration. This would help for better and real understanding of text mining. Proposed approach gives efficient output with the help of NLP. Open NLP tools are freely available. Tools included in the C# port are: examples are a sentence splitter, a tokenizer, a part-of-speech tagger, a chunker used to ("find non-recursive syntactic annotations such as noun phrase chunks"), a parser, and a name finder.The sample text input file from CORA dataset is shown in Figure 3. Open NLP tool is used for pre-processing. The Figure 4 shows the screenshot of Open NLP tool.

A. *Spilt*

Splitting is the term used for detecting the end of sentences. If we have input as a text paragraph in a string format, a simple way dividing paragraph into sentences. It uses syntax input. Split ('.') to obtain an array of strings. Extending this to input. Split ('.', '!', '?'), whenever punctuation mark is occurs, sentences are splitting. But this technique does not recognize that punctuation mark can appear in the middle of sentences too. Figure 5 shows a screenshot of sentence splitting.

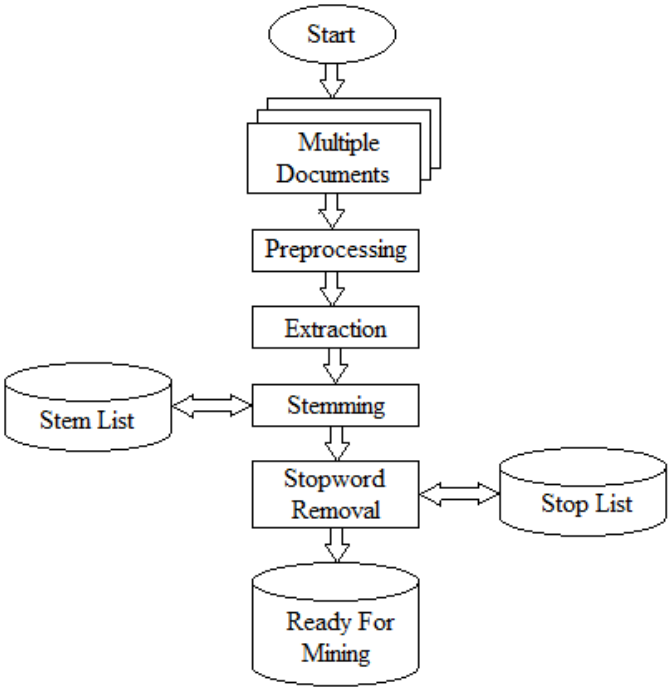
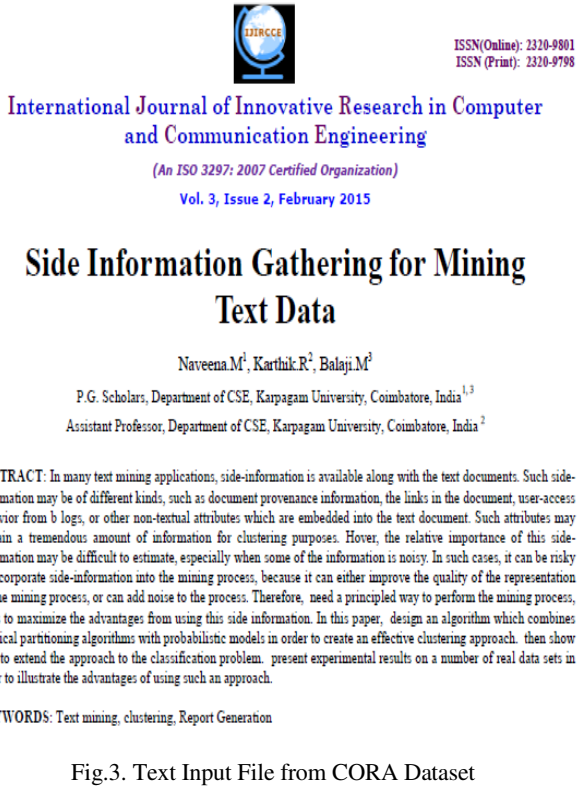


Fig.2. Work Flow of Proposed Work



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. 3, Issue 2, February 2015

Side Information Gathering for Mining Text Data

Naveena.M[1], Karthik.R[2], Balaji.M[3]

P.G. Scholars, Department of CSE, Karpagam University, Coimbatore, India[1,3]

Assistant Professor, Department of CSE, Karpagam University, Coimbatore, India[2]

ABSTRACT: In many text mining applications, side-information is available along with the text documents. Such side-information may be of different kinds, such as document provenance information, the links in the document, user-access behavior from b logs, or other non-textual attributes which are embedded into the text document. Such attributes may contain a tremendous amount of information for clustering purposes. Hover, the relative importance of this side-information may be difficult to estimate, especially when some of the information is noisy. In such cases, it can be risky to incorporate side-information into the mining process, because it can either improve the quality of the representation for the mining process, or can add noise to the process. Therefore, need a principled way to perform the mining process, so as to maximize the advantages from using this side information. In this paper, design an algorithm which combines classical partitioning algorithms with probabilistic models in order to create an effective clustering approach. then show how to extend the approach to the classification problem. present experimental results on a number of real data sets in order to illustrate the advantages of using such an approach.

KEYWORDS: Text mining, clustering, Report Generation

Fig.3. Text Input File from CORA Dataset

B. Tokenizing sentences:

Apply some NLP technique to it - part-of-speech tagging, or full parsing, perhaps. The first step in this process is to split the sentence into "tokens" - that is, words and punctuations.The "Tokenize" button in the tools example splits text in the top textbox into sentences, and then tokenizes each sentence. The output, in the lower textbox, places pipe characters between the tokens. Tokenization separated by bar symbol. Figure 6 shows a screen shot of sentence tokenization.

C. Part-of-speech (POS) tagging

Part-of-speech (POS) tagging use for providing a part of speech to each word in a sentence. Input is an array of tokens from the tokenization process, proposed shows word is noun or pronoun. Figure 7 shows part-of-speech tagging from tokenization array.

D. Chunking (Finding phrases)

The OpenNLPtool will group the tokens of a sentence into larger chunks. And each of chunk corresponding to a syntactic unit such as a noun phrase or a verb phrase.This is the next step on the way to full parsing, but it could also be useful in itself when meaning in a sentence larger than the individual words.A POS tagged set of tokens is used for chunking. Figure 8 shows a parsing of stream of tokens.

E. Name Finding

"Name finding" is the term used by the OpenNLP library for identification of author's name (as per our CORA dataset) - for example, people's names, locations, dates, date, day, time, money. The proposed approach works on the use of training data, and tokens.Figure 9 shows a screen shot of name finding.

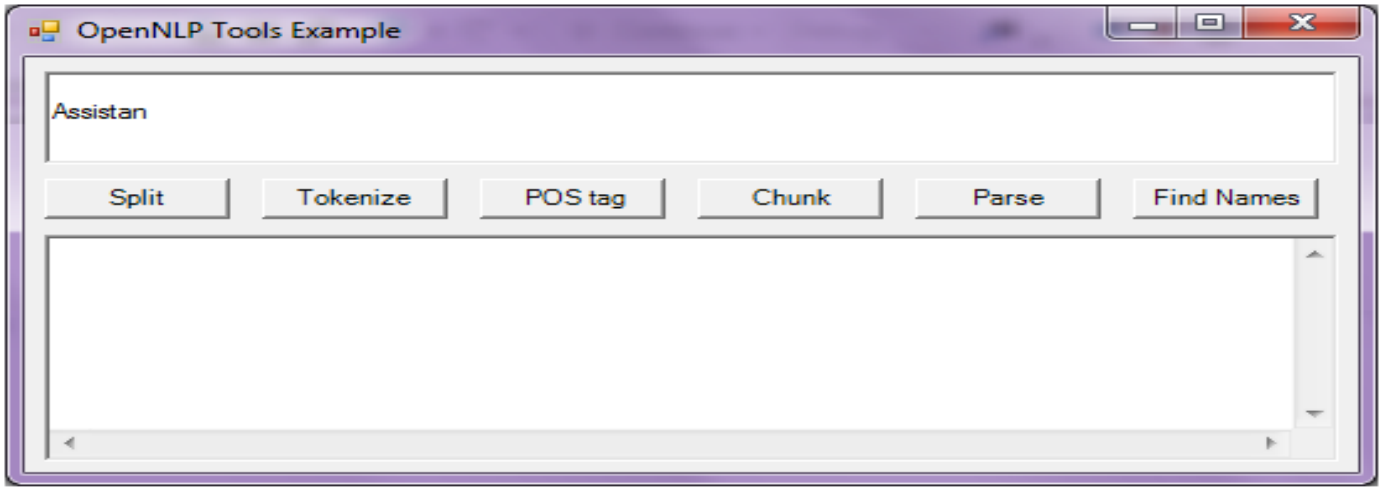


Fig.4. Screenshot of Open NLP tool

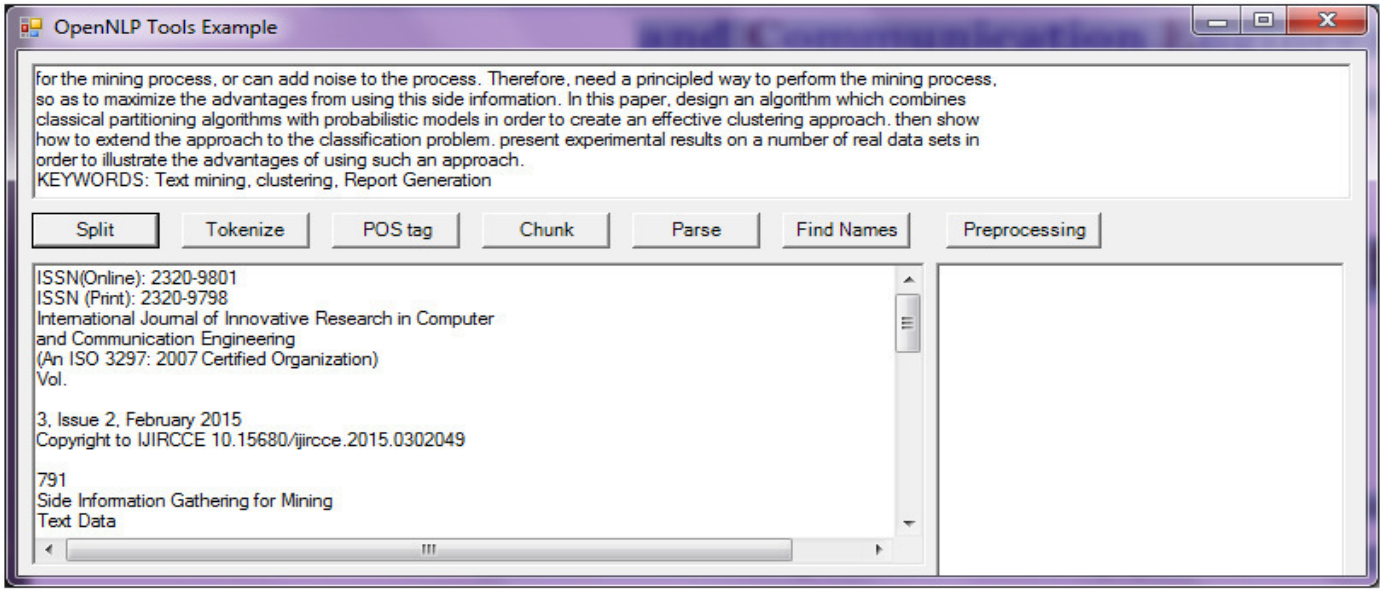


Fig.5. Splitting Sentences

Fig.6. Tokenization

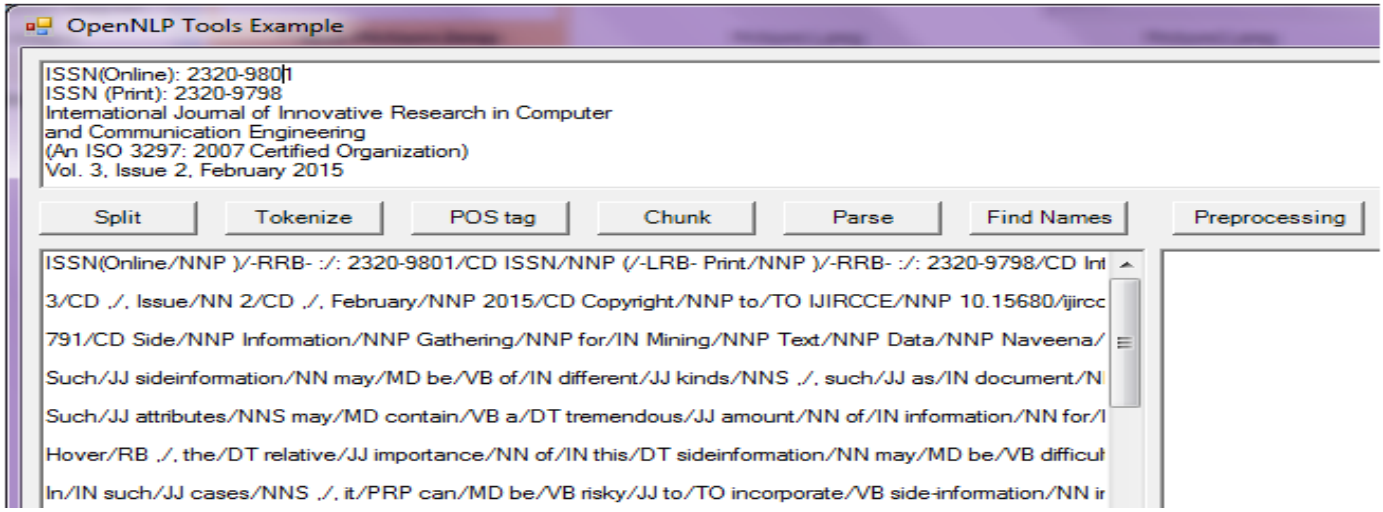


Fig.7. POS Tagging

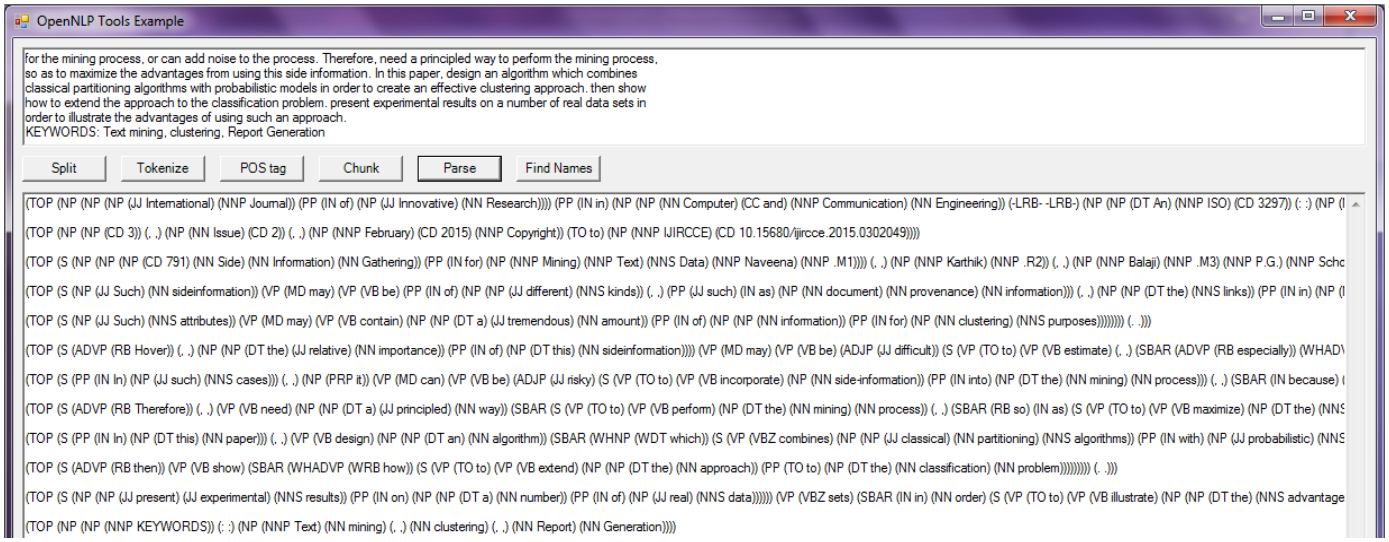


Fig.8. Parsing

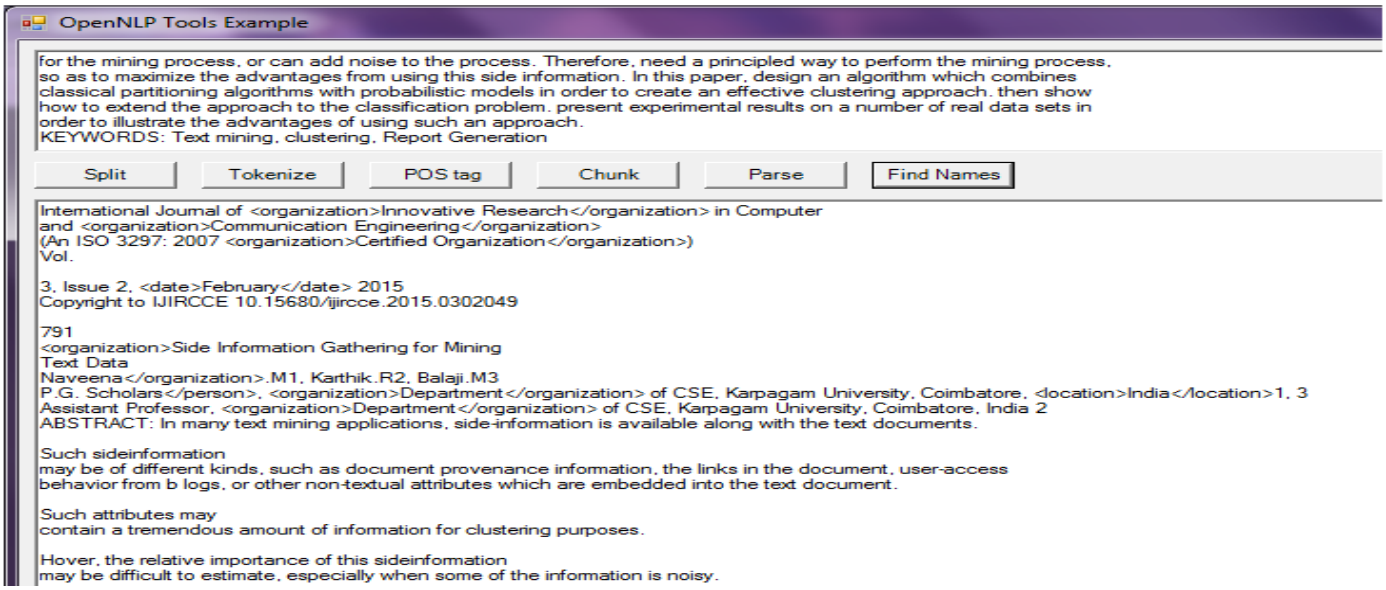


Fig.9. Find Naming

IV. Experimental Results

Figure 10 shows the graph for comparison of time required for mining different number of words. The time required for mining 2 words and 4 words are 55 ms and 360 ms, respectively. The time increased, it depends on the number of filtered words from the input document. It is natural that as the input increases the time required to mine input dataalso increases but not exponentially. Table I shows the time required for mining words with the count of filtered words.

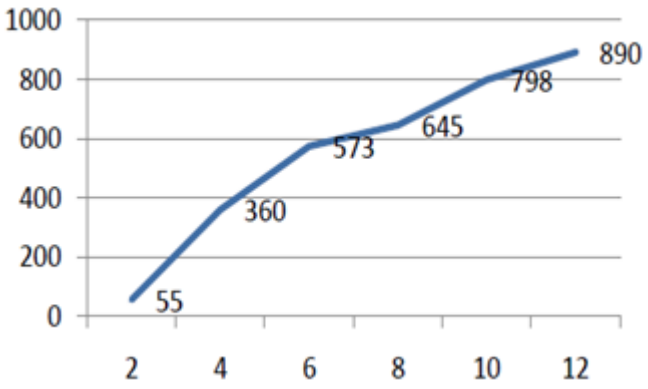


Fig.10. Comparison of Time Required for Mining Words

Table I: Time Required for Mining Words with the Count of Filtered Words for Different Number of Inputs

Input	*No. of Words*	*Filtered words*	*Time in ms*
File 1	2	292	55ms
File 2	4	1235	360ms
File 3	6	3128	573ms
File 4	8	4128	645ms
File 5	10	4450	798ms
File 6	12	4437	890ms

Proposed approach uses dictionary keywords comparing with input text file with dictionary words and matched keywords added to the process of side information generation. Figure 11 shows the graphical representation of word set and time required to generate side information.

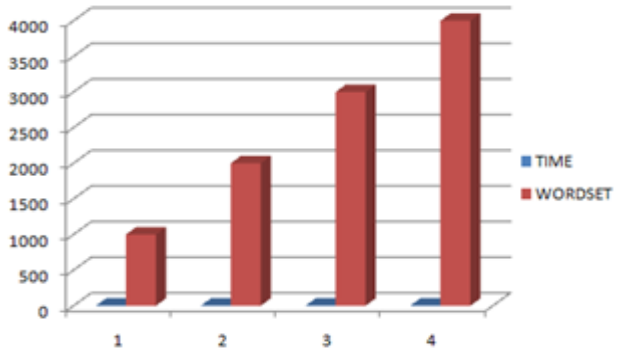


Fig.11. Comparison of Time Required to Generate Side Information for Different Wordset

Figure 12 shows the accuracy of input data. As the input data increases the accuracy remains constant i.e. number of words are more but also accuracy is high. The efficiency of side information generated is shown in Figure 13.

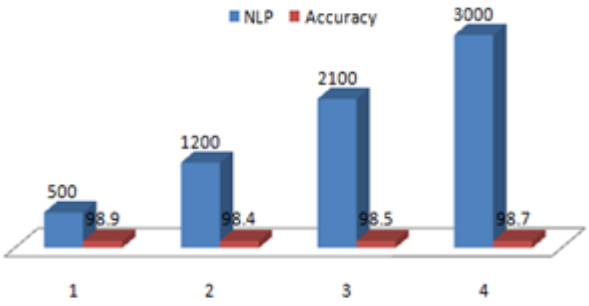


Fig.12. Accuracy of Input Data

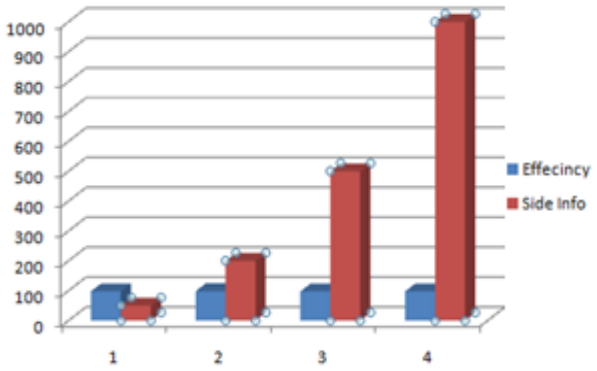


Fig.13. Efficiency of Side Information Generated

V. Conclusion and Future Scope

This paper presented the metadata based text mining for side information generation. The time required to mine input data depends on the size of input data. As the input data increases, the accuracy remains constant for generating the side information. Still, there is an improvement in designing a clustering based probabilistic approach for side information generation. Also, there is a scope for providing a security for side information generation and exploring the filter approaches. In future, a Security is provided for metadata based side information generation using Intrusion Detection System (IDS).

References

[1] S. Bhanuse, S. Kamble and S. Kakde, "Text Mining using Metadata for Generation of Side information", Procedia Computer Science, vol. 78, pp. 807-814, 2016.

[2] C. Aggarwal and P. Yu, "A framework for clustering massive text and categorical data streams", international Conference on Data. Mining, pp. 477–481, 2006.

[3] S. Guha, R. Rastogi, and K. Shim, "CURE: An efficient clustering algorithm for large databases", in Proc. ACM SIGMOD Conf.., New York, NY, USA, 1998, pp. 73-84, 1998.

[4] S. Guha, R. Rastogi, and K. Shim, "ROCK: A robust clustering algorithm for categorical attributes", info. Syst., vol. 25, no. 5, pp. 345-366, 2000.

[5] T. Liu, S. Liu, Z. Chen, and W.-Y. Ma, "An evaluation of feature selection for text clustering", in Proc. ICML Conf.., Washington, DC, USA, 2003, pp. 488-495.

[6] C. Aggarwal, Yuchen Zhao, and Philip S. Yu, "On the Use of Side Information for Mining Text Data", IEEE Transactions on knowledge and data engineering, vol. 26, ,no.6, pp. 1415-1429,2014.

[7] C. Aggarwal and H. Wang, "Managing and Mining Graph Data", New York, NY, USA: Springer, 2010.

[8] C. Aggarwal and C. Zhai, "A survey of text classification algorithms", in Mining Text Data. New York, NY, USA: Springer,2012.

[9] T. Yang, R. Jin, Y. Chi, and S. Zhu, "Combining link and content for community detection: A discriminative approach", in Proc. 15th ACM SIGKDD international conference on knowledge discovery and data mining, pp. 927-936, 2009.

[10] C. Aggarwal, Y. Zhao, and P. Yu, "On the Use of Side Information for Mining Text Data", IEEE Transactions on knowledge and data engineering vol. 26,no.6, pp. 1415-1429,2014.

[11] T. Zhang, R. Ramakrishnan, and M. Livny, "BIRCH: An efficient data clustering method for very large databases," in Proc., ACM SIGMOD Conf., New York, NY, USA, pp. 103–114, 1996.

[12] M. Khatri, S. Dhande "Implementation with text mining using classification", International Journal for Technological Research In Engineering, vol. 2, Issue 10, June-20.

[13] M. Steinbach, G. Karypis and V. Kumar, "A comparison of document clustering techniques," in Proc. Text Mining Workshop KDD, pp. 109-115, 2000.

[14] R. Feldman, J. Sanger "The Text Mining Handbook", Cambridge University Press, 2007.

[15] H. Mahgoub, and D. Rösner, "Mining association rules from unstructured documents," in Proc. 3rdInt. Conf. on Knowledge Mining, ICKM, Prague, Czech Republic, Aug. 25-27, pp.167-172, 2006.

[16] A. McCallum. "Bow: A Toolkit for Statistical Language Modeling, Text Retrieval", Classification and Clustering, 1996, http://www.cs.cmu.edu/~mccallum/bow/

[17] R. Angelova and S. Siersdorfer, "A neighborhood-based approach for clustering of linked document collections," in Proc. CIKM Conf., New York, NY, USA, pp. 778-779, 2006.

[18] A. Jain and R. Dubes, "Algorithms for Clustering Data", Englewood Cliffs, NJ, USA: Prentice-Hall, Inc., 1988.

[19] I. Dhillon. S. Mallela, and D. Modha, "Information-theoretic Co-clustering," In Proc. ACM KDD Conf., New York, NY, USA, pp. 89-98, 2003.

[20] A. Banerjee and S. Basu, "Topic models over text streams: A study of batch and online unsupervised learning," In Proc. SDM Conf., pp. 437-442, 2007.

[21] Y. Sun, J. Han, J. Gao, and Y. Yu, "Topic Model: Information network integrated topic modeling," In Proc. ICDM Conf., Miami, FL, USA, pp. 493-502, 2009.

[22] Ning Zhong, Yuefeng Li, and Sheng-Tang Wu," Effective Pattern Discovery for Text Mining", IEEE Transactions on Knowledge and Data Engineering, vol. 24, no. 1,2012.

[23] M. Franz, T. Ward, J. S. McCarley, and W. J. Zhu, "Unsupervised and supervised clustering for topic tracking", In Proc. ACM SIGIR Conf., New York, NY, USA, pp. 310-317,2001.