International Journal of Science and Research (IJSR)
ISSN (Online): 2319-7064
Index Copernicus Value (2013): 6.14 | Impact Factor (2014): 5.611

Document Classification Using Part of Speech in Text Mining

Sonam Tripathi[1], Tripti Sharma[2]

[1]Department of Computer Science and Engineering, Rungta College of Engineering and Technology, Kohka Kurud Road, Bhilai

[2]Department of Computer Science and Engineering, Rungta College of Engineering and Technology, Kohka Kurud Road, Bhilai

Abstract: ***Text mining is a practice that is used to find beneficial in arrangement from the large amount of data sets. Data mining has guidelines called as frequent pattern and association rule that is important for finding frequent patterns. Text Mining is the detection by computer of new, previously unidentified in arrangement by automatically mining in arrangement from dissimilar written resources. Text mining methods are the fundamental and permitting tools for efficient organization, triangulation, retrieval and summarization of large document quantity. The problem is more often than not decomposed into two sub problems. The first is to find those kind of item sets whose occurrence goes beyond a predefined threshold set in the database; those item sets are describe frequent or large item sets. The second problem is to produce involvement rules from those huge item sets with the restriction of minimal self-confidence. In this work, the text mining is done by dividing the given set of paragraphs into tokens and classifying them accordingly. The techniques are purely composed of sequential pattern mining, closed pattern mining & frequent pattern mining. Hence, the discovered patterns in the field of text mining cannot be used further or again. All frequently used short patterns are not useful here. In this work, an effective pattern taxonomy model & part of speech have been proposed to overcome and solve the problem of low frequency & misinterpretation.***

Keywords: Text mining; Association rule; Sequential pattern mining; Closed pattern mining; Frequent pattern mining.

1. Introduction

Text mining is a practice that is used to find beneficial in arrangement from large amount of data sets. Data mining has guidelines called as frequent pattern and association rule that is important for finding frequent patterns. Text Mining is the detection by computer of new, previously unidentified in arrangement, by automatically mining in arrangement from dissimilar written resources. Text mining methods are the fundamental and permitting tools for efficient organization, triangulation, retrieval and summarization of large document quantity. With more and more text, in arrangement are spreading around on Internet, text mining is growing in importance. Text clustering and text classification are two essential tasks in the field of text mining.

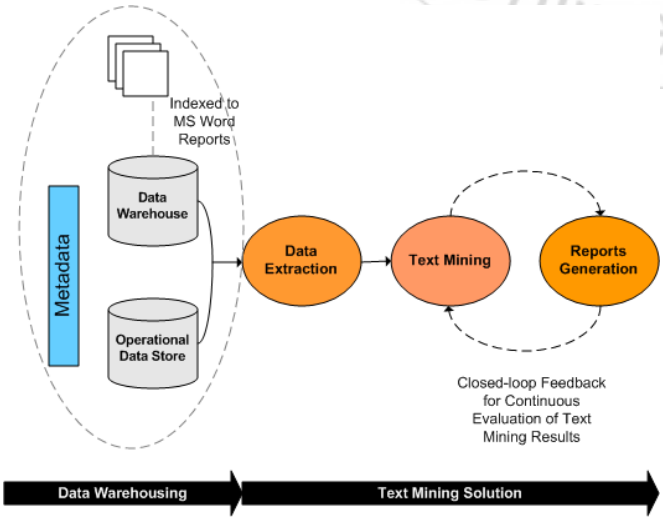

Figure 1: Process of Text Mining Block Diagram

Text Clustering is to find out the groups in arrangement from the text documents and cluster these documents into the most relevant groups. Text clustering clusters the document in an unsupervised way and there is no label or class in arrangement. Clustering methods have to determine the connections between the document and then based on these connections the documents are bunched. Given enormous volumes of documents, a good document clustering technique may organize those huge statistics of documents into meaningful groups, which allow further browsing and navigation of this quantity be much easier. A basic idea of text clustering is to find out which kind of documents have many words in common and place these kind of documents with the most words in communal into same group.

Text Classification is to establish the documents into predefined classes with meaningful labels. As text classification needs the facts about those predefined categories, it is applied in a supervised way.

Eighty percent of the in arrangement in the world is presently stored in amorphous textual arrangement. Although methods such as Natural Language Processing (NLP) can complete limited text analysis, there are presently no computer programs available to investigate and interpret text for the diverse in arrangement extraction needs. Thus text mining is a dynamic and unindustrialized area. The world is fast becoming in arrangement exhaustive, in which specialized in arrangement is being poised into a very large data sets. For example, Internet contains a large amount of online text documents, which quickly change and grow. It is nearly dreadful to manually organize such vast and rapidly evolving in arrangement. The necessity to extract useful and relevant in arrangement from such bulky data sets has led to an important requirement to develop computationally competent text mining algorithms. An example problem is to automatically dispense natural language text documents to predefined sets of categories grounded on their content. Other examples of problems involving large data sets

comprise searching for targeted in arrangement from technical citation databases (e.g. MEDLINE); search, filter and classify web pages by topic and routing relevant email to the suitable addresses.

Text mining is the involuntary and semi-automatic extraction of implicit, previously indefinite, and hypothetically useful in arrangement and patterns, from a large amount of amorphous textual data, such as natural-language texts. In text mining, every document is represented as a vector, whose dimension is almost the number of different keywords in it, which can be very large. One of the main contests in text mining is to categorise textual data with such elevated dimensionality. In adding up to high dimensionality, text-mining algorithms would also pact with word ambiguities such as pronouns, synonyms, and deafening data, spelling mistakes, abbreviations, acronyms and inadequately structured text. Text mining algorithms are of two kinds: Supervised learning and unsupervised learning. Support vector machines (SVMs) are a set of supervised learning approaches used for classification and reversion. Nonnegative matrix factorization is an unsupervised learning method.

2. Problem Identification

Large document corpus may afford a lot of useful information to people. But it is also a challenge to come across out the useful in sequence from huge number of documents. Especially with the detonate of knowledge around the cyber-world, corporate and organizations demand efficient and ineffectual ways to systematize the large document corpus and make later navigating and browsing to be converted into more easy, friendly and efficient. An obvious distinguishing of large document corpus is the enormous volumes of documents. It is almost impossible for a man to read through all the documents and come across out the relative for a specific topic. How to organize large document corpus is the problem we concern.

It was obtainable that the formal characterization of the problem of chronological pattern mining and its application to demining the web log. Given (i) a set of chronological records (called sequences) on behalf of a sequential database D; (ii) a smallest amount support threshold called min sup ξ ; and (iii) a set of k only one of its kind items or events I = {i1, i2, . . . , ik}. The problem of mining sequential patterns is that of pronouncement the set of all frequent progressions S in to the agreed sequence database D of items I at the given min sup.

In many real applications more than ever in dense data with long frequent patterns enumerating all potential 2L – 2 subsets of an L length pattern is infeasible. A reasonable solution is identifying a slighter envoy set of patterns from which all other frequent patterns can be consequent. Maximal recurrent patterns (MFP) form the smallest representative set of patterns to engender all frequent patterns. In scrupulous, the MFP are those patterns that are frequent but not an iota of their supersets are frequent. The problem of maximal recurrent patterns mining is finding all MFP in D with high opinion to σ.

The complexity of recurrent patterns mining from a large quantity of data is generating a huge number of patterns disappointing the smallest amount sustain threshold, especially when min_sup σ is precise low. This is because, all sub-pattern of a recurrent pattern are frequent as well. Consequently a long pattern contains a number of shorter frequent sub patterns. Assorted category of frequent patterns can be excavation from different kinds of data sets. In this make inquiries, we utilize item sets (sets of items) as a data set and the wished-for method is for frequent item set mining, that is, the mining of frequent from transactional data sets. However, it can be wholesale for additional kinds of frequent patterns.

The problem is more often than not decomposed into two sub problems.

1) One is to find those item sets whose occurrence goes beyond a predefined threshold in the database; those item sets are describe frequent or large item sets.
2) The second problem is to produce involvement rules from those huge item sets with the restriction of minimal self-confidence. [Goswami D.N. et al, 2010]

3. Related Work

Zhong N. et al (2012) presented an innovative and effective pattern discovery technique which includes the processes of pattern deploying and pattern evolving, to improve the effectiveness of using and updating discovered the patterns for finding relevant and interesting information. Substantial experiments on RCV1 data collection and TREC topics demonstrated that the proposed solution achieves encouraging performance.

Luepol Pipan maekaporn (2013), presented a novel pattern mining approach to RF. This approach mined patterns in both positive and negative feedback and then classified them into clusters to find user-specific patterns. They also proposed a novel pattern deploying method that effectively used the discovered patterns for improving the performance of searching relevant documents. Experiments are conducted on Reuters Corpus Volume 1 data collection (RCV1) and TREC filtering topics. The results shown that the proposed approach achieves promising performance comparing with state-of-the art term-based methods and pattern-based ones.

They also applied a novel pattern deploying the strategy to improve the performance of frequent patterns in text. They evaluated the proposed approach by using it to discover high-quality features in relevance feedback for improving the information filtering. Their results on RCV1 data collection and TREC filtering topics confirmed that the best improvements are obtained by our approach compared to state-of-the-art term-based methods and pattern-based ones.

Bhushan Inje, Ujawla Patil (2014) examined and investigated this fact with considering several states of art data mining methods that gives satisfactory results to improve the effectiveness of the pattern. Here they implemented the pattern detection method to solve problem of term-based methods and improved result which is helpful in information retrieval systems. Their proposal was also evaluated for several they'll distinguish domain, offering in

all cases, reliable taxonomies considering precision and recall along with F-measure. For the experiment, they used Reuters (RCV1) dataset and the results show that they improved the discovering pattern as compared to previous text mining methods. The results of the experiment setup show that the keyword-based methods not give better performance than pattern-based method. The results also indicated the removal of meaningless patterns not only reduces the cost of computation but also improved the effectiveness of the system.

Rupali Gangarde and V.L. Kolhe (2014) an effective pattern discovery technique is given which applies a pattern co-occurrence matrix to clean close sequential patterns. The Process of pattern deploying is applied with the co-occurrence and absolute support (PDCS) as deploying approach to overcome pattern misinterpretation problems and pattern evolving to overcome low frequency problem. They also applied a pattern co-occurrence matrix to clean close sequential patterns. This improved performance by using and updating discovered patterns and finding interesting and relevant information.

Mabroukeh N.R. and Ezeife C.I (2010) presented taxonomy of sequential pattern-mining techniques in the literature with theyb usage mining as an application. This article investigates these algorithms by introducing taxonomy for classifying sequential pattern-mining algorithms based on important key features supported by the techniques. This classification aims at enhancing understanding of sequential pattern-mining problems, current status of provided solutions, and direction of research in this area. This article also attempts to provide a comparative performance analysis of many of the key techniques and discusses theoretical aspects of the categories in the taxonomy.

Han J. et al (2007) provide a brief overview of the current status of frequent pattern mining and discuss a few promising research directions. They believe that frequent pattern mining research has substantially broadened the scope of data analysis and will have deep impact on data mining methodologies and applications in the long run. However, there are still some challenging research issues that need to be solved before frequent pattern mining can claim a cornerstone approach in data mining applications. They presented a brief overview of the current status and future directions of frequent pattern mining. With over a decade of extensive research, there have been hundreds of research publications and tremendous research, development and application activities in this domain. It was impossible for them to give a complete coverage on this topic with limited space and our limited knowledge.

4. Methodology

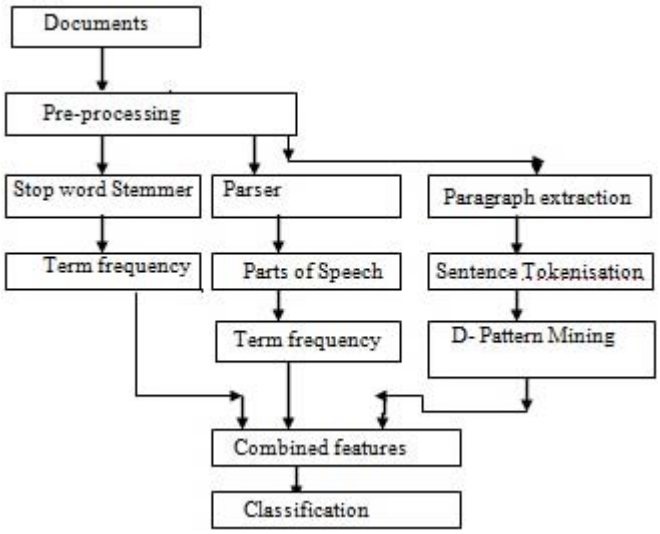

Figure 2: The Flowchart of the Methodology

Step By step Explanation of the Methodology

Step 1:- Selection of the Documents.

- Take the paragraph for testing in a random manner.

Step 2:-Pre processing of the Documents.

- Arrange the data in a prescribed manner or make a data set ready for processing.
- The data should be in a well arranged format, so that it could be accessible.

Step 3(a):- Stop word Stemmer

- To the default stop word, the lists connect in various ways with the stemming algorithm.

(b):- Term Frequency

- Term frequency, which is inverse document frequency is a numerical statistics
- It is intended to reflect how important a word is to a document in a collection of corpus. It is often used as a weighting factor in information retrieval & text mining.

Step 4 (a):- Parser

- In computer technology a parser is a program usually a part of compiler that receives input in the form of sequential source program instruction, interactive online command, mark-up text or some other defined interface & breaks them up into parts.
- The Parser receives the input from the document, processes it by breaking them into its constituent parts.

(b):- Part of speech

- The Program identifies the parts of speech of the paragraph or the sentences

(c):- Term Frequency

- Term frequency, which is inverse document frequency is a numerical statistics
- It is intended to reflect how important a word is to a document in a collection of corpus. It is often used as a weighting factor in information retrieval & text mining.

Step 5(a):- Paragraph Extraction

- Extract the keywords & the key phrases from any paragraph without changing the meaning of paragraph.

(b):- Sentence tokenization

- Sentence tokenization is a way to split text into token like paragraph, or individual words.
- Token the sentences into smaller sentences or even the individual words.

(c):- D-pattern mining

- Mining algorithm for discovering patterns from large datasets keyword text mining text classification pattern mining.

Step 6:- Combined features

- Combine all the features of the document
- This is a desired output.

Step 7:- Classification

- Classify the desired output.

5. Results and Discussions

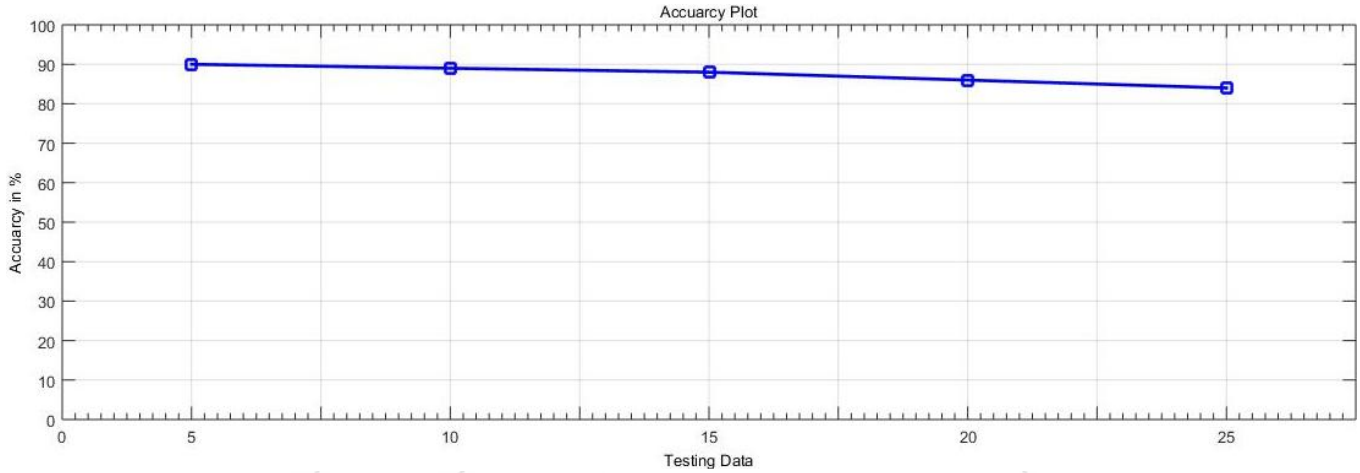

Figure 3: Comparison of the training data vs. accuracy rate

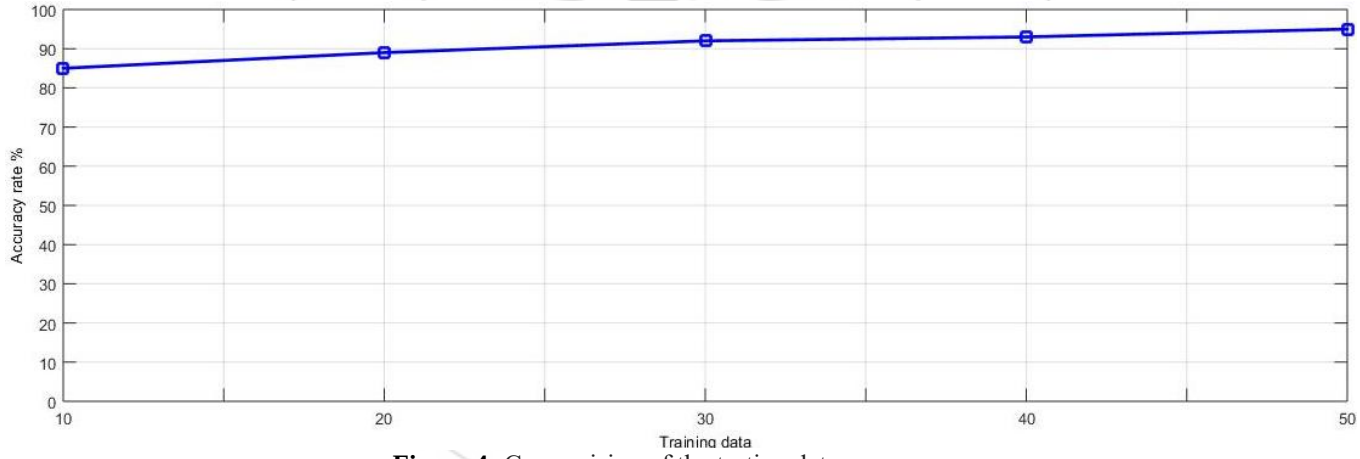

Figure 4: Comparision of the testing data vs accuracy

6. Conclusion and Future Work

In the last decade, several data mining techniques had been implemented for completing the various knowledgeable discovery tasks. These techniques are purely composed of sequential pattern mining, closed pattern mining & frequent pattern mining. Hence, these discovered patterns in the field of text mining cannot be used further or again. All frequently used short patterns are not useful here. In this work, an effective pattern taxonomy model & part of speech have been proposed to overcome and solve the problem of low frequency & misinterpretation problem.

Telescoping in tree projection can be applied to pattern-growth algorithms to reduce the tree size, where more than one item can be compressed into one node or edge. Using the Fibonacci sequence to partition or sample the search space may be useful for effective mining of very long sequences. Distributed mining of sequences can provide a way to handle scalability in very large sequence databases and long sequences. In the area of web usage mining. This can be applied to mine several web logs distributed on multiple servers.

References

[1] Anisha Radhakrishnan, Mathew Kurian, 2013, "Efficient Updating Of Discovered Patterns For Text Mining: A Survey", "IJCSNS International Journal Of Computer Science And Network Security, VOL.13 No.10", Pp 104-108.

[2] Anisha Radhakrishnan, Mathew Kurian, 2013, "Effective Pattern Matching Approach For Knowledge Discovery Application", "International Journal Of Advanced Research In Electronics And Communication

Engineering (IJARECE) Volume 2, Issue 2", Pp.224-226.
[3] BhushanInje, UjawlaPatil, 2014, "Operational Pattern Revealing Technique In Text Mining", IEEE Students' Conference On Electrical, Electronics And Computer Science.
[4] DiptiS.Charjan, Prof.MukeshA.Pund , 2013, "Pattern Discovery For Text Mining Using Pattern Taxonomy", "International Journal Of Engineering Trends And Technology (IJETT) – Volume 4 Issue 10", Pp. 4550-4554.
[5] Feng-Gang Li, Ying-Jia Sun, Zhi-Wei Ni, Yu Liang, Xue-Ming Mao, 2012, "The Utility Frequent Pattern Mining Based On Slide Window In Data Stream", "Fifth International Conference On Intelligent Computation Technology And Automation", Pp. 414-419.
[6] Goswami D.N., ChaturvediAnshu, Raghuvanshi C.S., 2010, "An Algorithm For Frequent Pattern Mining Based On Apriori", "(IJCSE) International Journal On Computer Science And Engineering Vol. 02, No.04", Pp.942-946.
[7] Jiawei Han · Hong Cheng · Dong Xin · Xifeng Yan, 2007, "Frequent Pattern Mining: Current Status And Future Directions", "Data Min Knowl Disc", Pp. 56-78.
[8] JyotiJadhav, LataRagha, Vijay Katkar, 2012, "Incremental Frequent Pattern Mining", "International Journal Of Engineering And Advanced Technology (IJEAT) ISSN: 2249 – 8958, Volume-1, Issue-6", Pp 222-226.
[9] K. Mythili, Professor, K. Yasodha,2012, "A Pattern Taxonomy Model With New Pattern Discovery Model For Text Mining", "International Journal Of Science And Applied Information Technology Volume 1, No.3", Pp. 88-92.
[10] K. Sasireka, G. Kiruthiga, And K. Raja, 2011, "A Survey About Various Data Structures For Mining Frequent Patterns From Large Databases", "International Journal Of Research And Reviews In Information Sciences (IJRRIS) Vol. 1, No. 3,", Pp. 85-88.
[11] Karam Gouda A, MosabHassaan A, Mohammed J. Zaki B,∗, 2010, "Karam Gouda A, MosabHassaan A, Mohammed J. Zaki B,∗" Journal Of Computer And System Sciences 76", Pp. 88-108.
[12] LuepolPipanmaekaporn, 2013, "Feature Discovery In Relevance Feedback Using Pattern Mining" IEEE, Pp. 301-312.
[13] Mr. Rahul Mishra, Ms.AbhaChoubey, 2012, "Discovery Of Frequent Patterns From Web Log Data By Using FP-Growth Algorithm For Web Usage Mining", "International Journal Of Advanced Research In Computer Science And Software Engineering", "Volume 2, Issue 9", Pp. 311-317.
[14] NingZhong, Yuefeng Li, And Sheng-Tang Wu , 2012 , " Effective Pattern Discovery For Text Mining", "IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 24, NO. 1", Pp 30-44.
[15] NIZAR R. MABROUKEH AndC. I. EZEIFE, 2010, "A Taxonomy Of Sequential Pattern Mining Algorithms", ACM Computing Surveys, Vol. 43, No. 1, Article 3,
[16] Norwati Mustapha, Mohammad-HosseinNadimi-Shahraki, Ali B Mamat, Md. Nasir B Sulaiman, 2005 – 2009, "A NUMERICAL METHOD FOR FREQUENT PATTERNS MINING", "Journal Of Theoretical And Applied Information Technology(JATIT)", PP. 92-97.

Author Profile

Ms. Sonam Tripathi received the B.E. degree from Chhattigarh Swami Vivekanand Technical University, Bhilai (C.G.) India in Computer Science & Engineering in the year 2012. She is currently pursuing M.Tech. Degree in Computer Science Engineering with specialization in Computer Science & Engineering from CSVTU Bhilai (C.G.), India. Her research area includes Data Mining and Text Mining etc.

Ms. Tripti Sharma is currently Assistant professor in Department of Computer science & Engineering RCET, Bhilai (C.G.) India. She completed her B.E and M.Tech. in Computer Science and Engineering Branch. Her research area includes Data mining, Image processing, Computer Network, AI & NN etc. She has published many Research Papers in various reputed National & International Journals, Conferences, and Seminars.