INTERNATIONAL RESEARCH JOURNAL IN ADVANCED ENGINEERING AND TECHNOLOGY (IRJAET) E - ISSN: 2454-4752 P - ISSN : 2454-4744
VOL 4 ISSUE 3 (2018) PAGES 3643 - 3649
RECEIVED : 25.04.2018 PUBLISHED : 31.05.2018
May 31, 2018

BRAINSTORMING PLATFORM USING MACHINE LEARNING TECHNIQUE

Qulsum Umer H Shaikh[1], Sachin Kumar[1], Aabriti Karki[1], P Aditya Rao[1]

School Of Engineering & Technology- Jain University, Bangalore

1. INTRODUCTION

In data mining text mining is one of the most important research areas in recent years. The rapid growth of evolution of technology the text documents usage can also increased. Such as, web pages, office documents and e-mails etc. Text mining can be automatically extracting information from different textual resources. It is multidisciplinary field, involving information retrieval, text analysis, and information extraction, clustering visualization, database technology, machine learning and data mining. The important aim of the text mining is to improve the textual database. All the paper publications and higher number of possible words and phrase types in the language, subtle and complex relationships between concepts in text. Information extraction is the task of automatically extracting structured information from unstructured and semi structured machine readable documents. Text mining is automatically extracting information from different textual resources. The goal of text mining is to discover previously unknown information. The challenges that arise due to unstructured text are large textual database. All publications are also in electronic form. Very high number of possible word and phrase types in the language, Complex and subtle relationships between concepts in text. A lot of research has been done to improve the quality of text representation and to develop high quality classifiers. Most of the machine learning methods as treats the text documents as bag of words. Information extraction (IE) is the task of automatically extracting structured information from unstructured and/or semi structured machine-readable documents. This activity concerns processing human language texts by means of natural language texts by means of natural language processing. The overall goal is to create a more easily machine readable text to process the sentences.

Text classification is an important part of text mining. Current research of text classification aims to improve the quality of text representation and develop high quality classifiers. Basically Text Mining tools can be divided into two parts that is Text analysis tools, Web Searching Tools. The text analysis tools are divided into four the document can be extracted using feature extraction, categorization, summarization, clustering. The clustering can be identified Hierarchical and binary relational clustering. Web searching tools can be analysed using Text Search Engine, net Question solution, web Crawler. Text classification process includes collection of data documents, data pre-processing, indexing, term weighing methods, classification algorithms and measure. Machine learning techniques have been actively explored for text classification.

An Internet forum, or message board, is an online discussion site where people can hold conversations in the form of posted messages. They differ from chat rooms in that messages are often longer than one line of text, and are at least temporarily archived. Also, depending on the access level of a user or the forum set-up, a posted message might need to be approved by a moderator before it becomes visible.

Forums have a specific set of jargon associated with them. Example a single conversation is called a "thread". A discussion forum is hierarchical or tree-like in structure: a forum can contain a number of sub forums, each of which may have several topics. Within a forum's topic, each new discussion started is called a thread, and can be replied to by as many people as so wish.

Depending on the forum's settings, users can be anonymous or have to register with the forum and then subsequently log in in order to post messages. On most forums, users do not have to log in to read existing messages

The problem of communication between students and teachers in an educational institution has become a major concern. This is because some students find it difficult to participate in classroom lectures because of their inability to socialize; also lecturers do not often have enough time to elaborate on the topics they have to teach for a particular class, hence, the decline in students' understanding of a given topic. The advent of computer mediated communication (CMC) has given rise to the development of online forum for effective communication. This journal highlights the structure and features of an online forum which makes it an effective communication tool between the lecturer and students of an institution.

The concept based analysis algorithm describes the concepts in the documents. This represents the semantic structures of the sentence and it is processed sequentially. Each concept in the present document is matched with other related concepts and also compared with previously processed documents concepts. Most of text mining techniques are based on word and /or phrase analysis of text. It is important to find term that contributes more semantic meaning to document this concept is known as concept based method. Only the importance of term within document is captured in statistical analysis of term based method. Only the importance of term within document is captured in statistical analysis of term based method. In concept based method the term which contributes to sentence semantic is analysed with respect to its importance at sentence and also document levels.

World Wide Web is an extremely large collection of information, i.e. beyond our imagination. It provides enough information according to user's need. Web is rising dreadfully as approximately 70 million pages are added daily. Knowledge Discovery on web data is referred as Web Mining. Web Structure Mining based on the analysis of patterns from hyperlink structure in the web. Like as Data Mining, Web Mining has four stages i.e. Data Collection, Preprocessing, Knowledge Discovery and Knowledge Analysis. This paper based on the first two stages Data collection and Preprocessing. Data collection is to collect the data required for analysis.

2. PROBLEM STATEMENT

Presently we were using plagiarism testing tools which treats the statement as single arrangement. if the same statement is changed into active or passive voice vice versa .it is unable to detect the copyright. So we need a method which can extract the exact meaning of sentence .It usually emphasizes the importance of grammatical division such as subject and predicate.

Plagiarism and copyright infringement overlap to a considerable extent, but they are not equivalent concepts, and many types of plagiarism do not constitute copyright infringement, which is defined by copyright law and may be adjudicated by courts. Plagiarism is not defined or punished by law, but

rather by institutions (including professional associations, educational institutions, and commercial entities)

when it comes to plagiarism, technology has been both a blessing and a curse. Though it has made it easier than ever to find and copy work from others without attribution, it's also made it easier to track and handle plagiarism when it happens.

With tools that can search billions of documents in seconds and can find matches only a few words in length, it might seem as if plagiarism would be as easily detected as finding information in Google. A matter of merely punching your query and going through the results.

Unfortunately, that isn't the case. Plagiarism detectors have a huge limitation and one that isn't likely to go away any time soon. That limitation is, simply put, that plagiarism detectors can't actually detect plagiarism and, instead, do something very different altogether.

3. THE EXISTING SYSTEM

This problem might seem a bit odd to those unfamiliar with the technology. After all, dishwashers wash dishes and car starters start cars, but plagiarism detectors don't actually detect plagiarism. Instead, what they actually detect is sections of identical text. Though there is a variety of techniques for doing this, the end results are pretty much always the same. A plagiarism detection service looks for matching strings of words between the document its looking at and the ones it has in its index. This is true for a local plagiarism checker. They all work on the same principle and basically function much like we would expect Google or another search engine to work, finding the words we want in other sources and providing the best results it can.

While this makes them powerful tools, doing the same comparison by hand would be impossible given all of the sources these tools can check, it does mean that it has some tremendous blind spots. However, those blind spots are only a problem if people aren't aware or don't believe that they are there. Then they become huge issues that can lead to both false positives and false negatives.

3.1 LIMITATION OF THE PRESENT SYSTEM

Since plagiarism detection tools can only detect copying, or more specifically similar phrases, there are two areas where they are particularly weak.

1. **Non-Verbatim Plagiarism:** Plagiarism that involves the rewriting, translating or otherwise redrafting the text can't be detected. This can be difficult to get away with as most plagiarism detectors are extremely sensitive, but since plagiarism detectors don't analyze the content of the work, just the words, it can't see if you lifted the idea or information if you didn't also lift the words. This is a common problem in academia, which treats this kind of plagiarism equally as seriously as verbatim plagiarism.

Common Phrasing/Attributed Use: Second, though many plagiarism checkers will make In short, plagiarism detection tools are just machines and they can make mistakes. However, that is true with any tool as, for example, you don't discard Microsoft Word because you can make a typo.

Also, like any other tools, plagiarism checkers are useless without humans to use them intelligently, which is the biggest problem such tools have an attempt to separate out attributed use, given the variety of attribution styles it isn't always possible. Also, given how common some phrases are in the English language, many plagiarism checkers will report matches that are actually just coincidence.

4. THE PROPOSED SYSTEM

Data preprocessing is an important and critical step in the data mining process, and it has a huge impact on the success of a data mining project. The purpose of data preprocessing is to cleanse the dirty/noise data, extract and merge the data from different sources, and then transform and convert the data into a proper format. Data preprocessing has been studied extensively in the past decade , and many commercial products such as Informatical and Data Joiner have been applied successfully in many applications. Most of the studies and commercial systems focus on data cleaning, extraction, and merging, even though some provide limited transformation capability, but they cannot meet the requirements of a lot of complicated data mining tasks.

A typical data set in data mining application tends to be high dimensional (hundreds even thousands of feature variables) with both numerical and symbolic type and has millions of tuples. Many actual applications, such as telephone billing, text categorization, and supermarket transactions, may collect hundreds to thousands of feature variables. Nonetheless, not all of the feature variables inherent in these applications are useful for sophisticated data analysis, for example, for data mining. One reason for this phenomenon is because most of the time, the data are collected without "mining" in mind. In addition, the existence of numeric data and the primitive symbolic values of symbolic attributes create a huge data space determined by the numeric data and primitive symbolic values. In order to mine the knowledge pattern from the data efficiently, it is essential to reduce the data set before the mining algorithm can be mined.

There are two directions to reduce the data set. One is to reduce the dimensions (attributes) of the data set by eliminating all those unnecessary attributes, and the other is to reduce the number of the tuples in the data set by discretizing the numeric attributes and generating the symbolic attribute values to high-level concept; a lot of tuples will be combined into one after the discretization and generalization, thus reducing the data tuples in the data set.

5. TERM FREQUENCY OCCURRENCE

World Server uses a simple and efficient word counting scheme. During the scoping process World Server breaks an asset into segments and then runs each segment through the word breaking process described above. After a list of word and number elements is generated, those words and numbers are counted using the following rules:

- Each word is counted as one point to be added to the total scoped value.

- For Chinese and Japanese, World Server has a special way to count words. Each character is considered a word. For these languages we are, effectively, counting characters. When a user sees "Words" in the World Server UI (for example, in scoping) for Chinese and Japanese source languages it actually means "Characters". If a content is a mixture of Chinese or Japanese and Latin-based languages, the appropriate word counting scheme is used for each language. For example, "World Server " is counted as 6 words.
- In Korean, word counting is based on white spaces, not characters.
- Numbers are ignored unless a segment has no words.
- If a segment has no words but has at least one number the whole segment is counted as one word count.
- Wordless and numberless segments do not contribute to the scoped result.

FORMULA

Key value = total number of keywords/number of occurrence ×100

A typical data set in data mining application tends to be high dimensional (hundreds even thousands of feature variables) with both numerical and symbolic type and has millions of tuples. Many actual applications, such as telephone billing, text categorization, and supermarket transactions, may collect hundreds to thousands of feature variables. Nonetheless, not all of the feature variables inherent in these applications are useful for sophisticated data analysis, for example, for data mining. One reason for this phenomenon is because most of the time, the data are collected without “mining” in mind. In addition, the existence of numeric data and the primitive symbolic values of symbolic attributes create a huge data space determined by the numeric data and primitive symbolic values.

CONCLUSION

This algorithm can make full use of the internal characteristics of key documents, overcome single word frequency weighting factor weights and the problem of insufficient ability of text characteristic in the classic TF-IDF algorithm, improving the precision of selection of key and text clustering accuracy. If scenario requires, more weighting factor should be introduced to the TF-IDF algorithm, time complexity of algorithm also should be avoid increase rapidly. Data preprocessing is an important and critical step in the data mining process, and it has a huge impact on the success of a data mining project. The purpose of data preprocessing is to cleanse the dirty/noise data, extract and merge the data from different sources, and then transform and convert the data into a proper format. Data preprocessing has been studied extensively in the past decade , and many commercial products such as Informatical and Data Joiner have been applied successfully in many applications. Most of the studies and commercial systems focus on data cleaning, extraction, and merging, even though some provide limited transformation capability, but they cannot meet the requirements of a lot of complicated data mining tasks.

REFERENCES

[1] An Efficient Concept-based Mining Model for enhancing Text Clustering, Shady Shehata, Fakhri Karray, Mohamed S.Kamel, IEEE Transactions on knowledge and Data Engineering, vol 22, no 10, October 2010.

[2] K.J. Cios, W. Pedrycz, and R.W. Swiniarski, Data Mining Methods for Knowledge Discovery. Kluwer Academic Publishers, 1998.

[3] B. Frakes and R. Baeza-Yates, Information Retrieval: Data Structures and Algorithms. Prentice Hall, 1992.

[4] K. Aas and L. Eikvil, "Text Categorisation: A Survey," Technical Report 941, Norwegian Computing Center, June 1999.

[5] G. Salton, A. Wong, and C.S. Yang, "A Vector Space Model for Automatic Indexing," Comm. ACM, vol. 18, no. 11, pp. 112- 117,1975.

[6] G. Salton and M.J. McGill, Introduction to Modern Information Retrieval. McGraw-Hill, 1983.

[7] U.Y. Nahm and R.J. Mooney, "A Mutually Beneficial Integration of Data Mining and Information Extraction," Proc. 17th Nat‟l Conf. Artificial Intelligence (AAAI „00), pp. 627-632, 2s000.

[8] L. Talavera and J. Bejar, "Generality-Based Conceptual Clustering with Probabilistic Concepts," IEEE Trans. Pattern Analysis and Machine Intelligence, vol. 23, no. 2, pp. 196-206, Feb. 2001.

[9] H. Jin, M.-L. Wong, and K.S. Leung, "Scalable Model-Based Clustering for Large Databases Based on Data Summarization," IEEE Trans. Pattern Analysis and Machine Intelligence, vol. 27, no. 11, pp. 1710-1719, Nov. 2005.

[10] T. Hofmann, "The Cluster-Abstraction Model: Unsupervised Learning of Topic Hierarchies from Text Data," Proc. 16th Int‟l Joint Conf. Artificial Intelligence (IJCAI „99), pp. 682-687, 1999.

[11] Eugene Agichtein , Silviu Cucerzan ,"Predicting Accuracy of Extracting Information from Unstructured Text Collections", CIKM'05, October 31-November 5, 2005.

[12] Eugene Agichtein, "Scaling Information Extraction to Large Document Collections", IEEE Computer Society Technical Committee on Data Engineering, 2005.

[13] Raymond J. Mooney and Razvan Bunescu, "Mining Knowledge from Text Using Information Extraction", SIGKDD Explorations, 2005. S.Brindha et al, International Journal of Computer Science and Mobile Computing, Vol.5 Issue.9, September- 2016, pg. 112-116 © 2016, IJCSMC All Rights Reserved 116

[14] Yung-Shen Lin, Jung-Yi Jiang, and Shie-Jue Lee, "A Similarity Measure for Text Classification and Clustering", IEEE Transactions on Knowledge and Data Engineering, 2013.

[15] Shady Shehata,Fakhri Karray, Mohamed S. Kamel, "An Efficient Concept-Based Mining Model for Enhancing Text Clustering", IEEE Transactions on Knowledge and Data Engineering, , 2010.

[16] Helena Ahonen and Oskari Heinonen "Applying Data Mining Techniques for Descriptive Phrase Extraction in Digital Document Collections"Published in the Proceedings of ADL"98, April 22-24, 1998 in Santa Barbara, California,USA.

[17] X. Li and B. Liu. Learning to classify texts using positive and unlabeled data. In IJCAI"03, pages 587–594, 2003.