Review Article Impact Factor: 4.226 ISSN: 2319-507X
Monika, IJPRET, 2015; Volume 3 (9): 796-803 IJPRET

INTERNATIONAL JOURNAL OF PURE AND APPLIED RESEARCH IN ENGINEERING AND TECHNOLOGY

A PATH FOR HORIZING YOUR INNOVATIVE WORK

REVIEW ON IMPLEMENTATION OF TEXT MINING WITH AUXILIARY INFORMATION USING CLASSIFICATION

MONIKA[1], PROF. S. S. DHANDE[2]

1. Student in Department of Computer Science & Engineering, Sipna College of Engineering and Technology, Amravati.
2. Associate Professor, Department of Computer Science and Engineering, Sipna College of Engineering and Technology, Amravati.

Accepted Date: 05/03/2015; Published Date: 01/05/2015

Abstract: Text mining is the analysis of data contained in natural language text. In many text mining applications such as web, social networking and other digital collections, etc side information is available along with text data. Such side information may be of different types such as document location, time, date, owner, last modified date, the links in the documents, user access behavior from web logs, etc. Side information may be useful for clustering and classification and may be curse sometimes. If the side information is noise free then it can help to cluster the text document with excellent results and if the side information is noisy then it can be risky for mining process. In such cases, it actually worsens the quality of clustering process. Therefore, we need a principled way to perform the mining process, so as to maximize the advantages of using this side information. Here we are going to perform mining on text and side information of the document with the help of natural language processing. In this project we will implement text mining with classification.

Keywords: Data mining, Natural language processing, Text mining, Classification

Corresponding Author: MS. MONIKA

Access Online On:

www.ijpret.com

How to Cite This Article:

Monika, IJPRET, 2015; Volume 3 (9): 796-803

PAPER-QR CODE

INTRODUCTION

Data mining is used to discern the large amount of information from various repositories. It can be useful for knowledge discovery and feature extraction from large amount of data. For example Google have been retrieved millions of data from various repositories. Text mining is the process of extracting some useful information from different structured and unstructured documents. Nowadays, most of the information is available in the form of text. The text is represented by following format such as word, phrase, term, pattern, concept, paragraph, sentence, and document. Data mining is also known as Knowledge Discovery in Data (KDD) [11]. Basically there are different types related to data mining, they are: text mining, web mining, multimedia mining, object mining and spatial data mining.

Text mining can help an organization originate potentially valuable business insights from text-based content such as word documents, email and postings on social media streams like Twitter, LinkedIn, etc. Text mining is the analysis of data contained in natural language text. Text mining works by transposing words and phrases in unstructured data into numerical values which can then be linked with structured data in a database and analyzed with traditional data mining techniques.

(Yoon and Lee, 2008) defined TM as "the process patterns or knowledge from unstructured branch of DM. The main aim of TM is analyzing and classifying a number of unstructured textual data and to discover the knowledge"[12].

In many application domains, a tremendous amount of side information is also associated along with the documents. This is because text documents typically occur in the context of a variety of applications in which there may be a large amount of other kinds of database attributes or meta information which may be useful to the clustering process. Some examples of such side-information are as follows:

• In an application in which we track user access behavior of web documents, the user-access behavior may be captured in the form of web logs. For each document, the meta-information may correspond to the browsing behavior of the different users. Such logs can be used to enhance the quality of the mining process in a way which is more meaningful to the user.

• Many text documents contain links among them, which can also be treated as attributes. Such links contain a lot of useful information for mining purposes.

• Many web documents have meta-data associated with them which correspond to different kinds of attributes such as the provenance or other information about the origin of the

document. In other cases, data such as ownership, location, or even temporal information may be informative for mining purposes. In a number of network and user-sharing applications, documents may be associated with user-tags, which may also be quite informative [1]

It uses techniques from information retrieval, information extraction as well as natural language processing (NLP) and connects them with the algorithms and methods of KDD, data mining, machine learning and statistics. Text mining can be also defined similar to data mining, information extraction and knowledge discovery process mode [10] [13].

The traditional NLP approach is: extract from the sentence a rich set of hand-designed features which are then fed to a standard classification algorithm [14].

LITERATURE REVIEW

Text Mining is the process of extracting interesting information or knowledge or patterns from the unstructured text that are from different sources. As the text is in unstructured form, it is quite difficult to deal with it. Finding nuggets of interesting information from the natural language text is the purpose of text mining. The Text Mining Process is:

Stage I: Preprocessing Text:

Mining from a preprocessed text is easy as compare to natural languages documents. So, preprocessing of documents that are from different sources is an important task during text mining process before applying any text mining technique. In Text mining, the selection of characteristics and also the influence of domain knowledge and domain-specific procedures play an important role [15].

Natural Language Processing (NLP): The general goal of NLP is to achieve a better understanding of natural language by use of computers. It employs simple and durable techniques for the fast processing of text. In addition, linguistic analysis techniques are used among other things for the processing of text[7][8].

In order to obtain all words that are used in a given text, a tokenization process is required, i.e. a text document is split into a stream of words by removing all punctuation marks and by replacing tabs and other non-text characters by single white spaces. The set of different words obtained by merging all text documents of a collection is called the dictionary of a document collection (bag-of-words representation). In order to reduce the size of the dictionary filtering and lemmatization or stemming methods can be adopted.

Filtering methods remove words like articles, conjunctions, prepositions from the dictionary and the same is used for the documents. Lemmatization methods try to map verb forms to the infinite tense and nouns to the singular form. Since this tagging process is usually quite time consuming and still error-prone, in practice frequently stemming methods are applied. Stemming methods try to build the basic forms of words, i.e. strip the plural 's' from nouns, the 'ing' from verbs, or other affixes. A linguistic preprocessing can be used to enhance the available information about terms. They perform the following methods:

(a) Part-of-speech tagging (POS) aims at labeling each word with a unique tag that indicates its syntactic role, e.g. plural noun, adverb.[12]

(b) Text chunking aims at grouping adjacent words in a sentence. Chunking, also called shallow parsing, aims at labeling segments of a sentence with syntactic constituents such as noun or verb phrase (NP or VP). Each word is assigned only one unique tag, often encoded as a begin-chunk (e.g. B-NP) or inside-chunk tag (e.g. INP).[18]

(c) Language Models, traditionally estimates the probability of the next word being w in a sequence. We consider a different setting: predict whether the given sequence exists in nature, or not, following the methodology of (Okanohara & Tsujii, 2007). This is achieved by labeling real texts as positive examples, and generating "fake" negative text[2]

(d) Semantically Related Words ("Synonyms") This is the task of predicting whether two words are semantically related (synonyms, holonyms, hypernyms...) which is measured using the Word Net database(http://wordnet.princeton.edu) as ground truth[3]

(e)Word Sense Disambiguation (WSD) tries to resolve the ambiguity in the meaning of single words or phrases.

(f)A Parsing produces a full parse tree of a sentence[9][10].

Stage II- Text Mining Technique is applied: This is an important stage in which the selected algorithm is applied on text in order to process the text. The algorithm such as clustering, classification, summarization, information extractions or visualizations could be used.

Stage III Analysis of Text:

Here the outputs are analyzed for discovering the knowledge. Various tools such as link discovery tool can be used or the outputs can be visualized so that the users could navigate through in order to achieve the perspective [15].

ANALYSIS OF PROBLEM

Clustering is the most common form of unsupervised learning and this is the major difference between clustering and classification. No supervision means that there is no human expert who has assigned documents to classes [16].

Clustering is sometimes incorrectly referred to as automatic classification; however this is incorrect, since the clusters found are not known prior to processing whereas in case of classification the classes are predefined [5].

In clustering, it is the distribution and the nature of data that will determine cluster membership, classes from a so called training set that is a set of data correctly labeled by hand and then replicates the learnt behavior on unlabeled data.

We are going to implement the text mining with classification for the most efficient text based classification. In clustering we will be able to get the output but not efficient one. So by using classification we will be able to obtain the efficient output.

For clustering we are going to use k-means algorithm because it is the simplest algorithm which uses unsupervised learning method to solve known clustering issues. It works really well with larger datasets.

Semantic score will be used which would be a count value of the number of elements matching in each cluster with the number of elements in input. This will provide the classified cluster which matches the input query.

PROPOSED WORK

1.Collection of datasets for data mining- In which we would be selecting various datasets and finding the best out of them for text mining with side information. This may be include methods like crawling, filtering, etc. It is recommended to accept big dataset because it works best on larger datasets.

2. Preprocessing data with natural language processing-In which we would be applying Natural language processing techniques like parts of speech tagging and chunking to find only the action words from the given text datasets. The field of Natural Language Processing (NLP) aims to convert human language into a formal representation that is easy for computers to manipulate. The general goal of NLP is to achieve a better understanding of natural language by

use of computers. It employs simple and durable techniques for the fast processing of text. In addition, linguistic analysis techniques are used among other things for the processing of text.

3. Development of mining algorithm with Natural Language Processing-for demonstration of text mining approach, we would developing the mining algorithm like k-means to extract data, and then produce the results at the output.

4. Combination of mining with natural language processing- In this all the above work would be combined in order to demonstrate our algorithm. This would demonstrate the use of NLP in text mining and obtaining the optimized outputs.

5.Result evaluation and optimization-In this module results would be evaluated and optimization would be performed to get the optimal outputs, comparison of results with and without the NLP algorithm would be done in order to get comparative analysis.

CONCLUSION

In Data mining is the most emerging field. Nowadays, almost all the work are done through electronic devices such as mobiles, computers, tablets, etc. In the same way internet is the most popular part of human life. Information available on internet is made available to user through various sources and so the information is unstructured. We as a user wants a structured data for our use. This structured data can be made available to us through mining techniques. In this project we are trying to implement the text mining approach that can be applied to any document and with the help of NLP and classification algorithm the side information is obtained in the documents.

REFERENCES

1. On the Use of Side Information for Mining Text Data Charu C. Aggarwal, *Fellow, IEEE* Yuchen Zhao, and Philip S. Yu, *Fellow, IEEE*

2. A Unified Architecture for Natural Language Processing: Deep Neural Networks with Multitask Learning Ronan Collobert Jason Weston NEC Labs America, 4 Independence Way, Princeton, NJ 08540 USA

3. Natural Language Processing (Almost) from Scratch by Ronan Collobert ,Jason Weston, L´eon Bottou, Michael Karlen, Koray Kavu kcuoglu, Pavel Kuksa NEC Laboratories America 4 Independence Way Princeton, NJ

4. Document Clustering by PankajJajoo, Indian Institute of TechnologyKharagpur2008Under the Guidance of Prof. Sudeshna Sarkar *Professor*

5. A SURVEY OF TEXT CLUSTERINGALGORITHMS by Charu C. Aggarwal *IBM T. J.* Watson Research Center Yorktown Heights, NY**,** Cheng Xiang Zhai University of Illinois at Urbana-Champaign Urbana.

6. A Survey on Improving the Clustering Performance in Text Mining for Efficient Information Retrieval S.Saranya#1, R.Munieswari*2 M.E Scholar, Department of Computer Science & Engineering, Kumaraguru College of Technology, Coimbatore, Tamilnadu, INDIA

7. NLP (Natural Language Processing) for NLP (Natural Language Programming) Rada Mihalcea Computer Science Department, University of North, Hugo Liu, and Henry Media Arts and Sciences, Massachusetts Institute of Technology

8. Natural Language Processing Gobinda G. Chowdhury Dept. of Computer and Information Sciences University of Strathclyde, Glasgow G11XH,UK

9. Bondale, N.; Maloor, P.; Vaidyanathan, A.; Sengupta,S. &Rao, P.V.S. (1999).Extraction of information fromopen-ended questionnaires using natural language processing techniques. Computer Science and Informatics, 29,15-22

10. Centre for Language Technology (2000). EAGLES-ll Information Page: Evaluation of NLP Systems. [Online]Available: http://www.cst.ku.dk/projects/eagles2.html

11. Revathi.T, Sumathi.P (2013),” A Survey on Data Mining using Clustering Techniques”, International Journal of Scientific & Engineering Research Volume 4, Issue 1.

12. Yoon Y. and Lee G. (2008). "Text Categorization Based on Boosting Association Rules", IEEE International Conference on Semantic Computing, (pp.136-143).

13. Natural Language Processing Laboratory, University of Massachusetts.

14. Black, W.J.; Rinaldi, F. & Mc Naught, J. (2000). Natural language processing in Java: applications in education and knowledge management. Proceedings of the Second International Conference on the Practical Application of Java.12-14 April 2000, Manchester. Practical Application Company: Blackpool. pp. 157-70

15. Text Mining Techniques-A survey, Divya Nasa, *USICT, GGSIPU,* ijarcsse, Volume 2. Issue 4, April 2012, pg.51-54

16. H. Schutze and C. Silverstein, "Projections for efficient document clustering," in *Proc. ACM SIGIR Conf.*, New York, NY, USA, 1997,pp. 74–81.

17. Classification Problem in Text mining, Jijy George, Sandhya N., Suja George, IJIRAE, Volume 1, Issue 8 (September 2014), pg.333-341

18. International Journal of Scientific and Research Publications, Volume 3, Issue 6, June 2013.ISSN 2250-3150.Machine Learning Algorithms for Opinion Mining and Sentiment Classification Jayashri Khairnar, Mayura Kinikar.