Classification of Text Document Based on Headword Extraction Algorithm using Text Mining

Kavya Jain[1] Dr.Siddharth Choubey[2]

[1,2]Department of Computer Science and Engineering

[1,2]Shri Shankaracharya Technical Campus, Junwani, Bhilai

Abstract— Text mining is a practice that is utilized to find advantageous information from large amount of data sets. Data mining has guidelines known as frequent pattern and association rule that is essential for finding frequent patterns. mining the semantic information from free text document provides the enabling technology for a host to identify the class to which the text document belongs. The NER (Noun Entity Region) has been used to identify the noun keywords using classifier uniquely viz 3 – Class classifier, 4 – class classifier and 7 – class classifier. By using the concept of Parsing and NER, the text document has been classified to the predefined class to which the given text document belongs by merging the MP – I (List of noun) and MP – II (List of verb) and matched it with the stored headwords to conclude which class the document belongs to. The use of parser to convert the text document to parse tree increased the accuracy of the work.

Key words: Text Mining, Noun Entity Recognizer, 3 – Class Classifier, 4 – Class Classifier, 7 – Class Classifier

I. Introduction

Text mining is a practice that is utilized to find advantageous information from large amount of data sets. Data mining has guidelines known as frequent pattern and association rule that is essential for finding frequent patterns. Text Mining is the recognition by computer of new, previously unidentified information, by automatically mining information from different written resources. Text mining techniques are the fundamental and permitting tools for efficient organization, triangulation, retrieval and summarization of large file quantity. With more and more text, information are distribution around on Internet, text mining is rising in importance. Text clustering and text classification are two important tasks in the field of text mining.

Supervised learning is a technique in which the algorithm customs predictor and target attribute value couples to learn the predictor and target value relation. Support vector machine is a supervised learning process for making a decision function with an exercise dataset. The training data contain pair of predictor and target values. Each predictor value is considered with a target value.

Unsupervised learning is a technique in which the algorithm practices only the predictor attribute values. There are no object attribute values and the learning chore is to add some understanding of pertinent structure patterns in the data. Each row in a data set signifies a point in n-dimensional space and unsupervised learning algorithms examines the relationship among these numerous points in n-dimensional space.

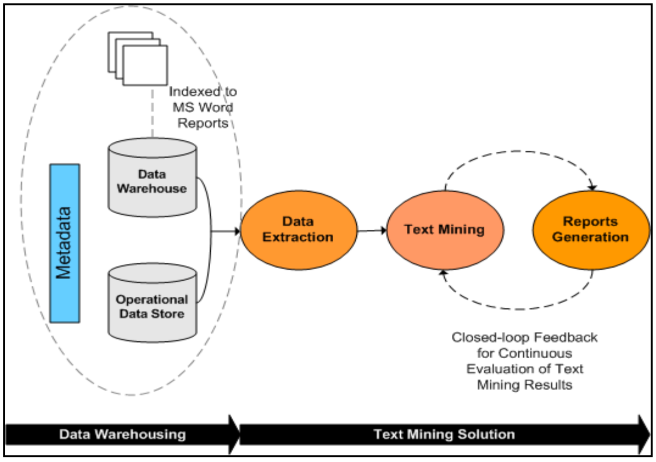


Fig. 1: Process of Text Mining Block Diagram

There are approximately five main methods in the text mining process: document retrieval, data extraction, data pre-processing (cleansing), data examination, and data visualization.

A. *Document Retrieval*

Information (or document) retrieval is a discipline apprehensive with the organizing, storage, searching, and retrieval of bibliographic information. Salton and McGill (1983) introduce the thought of the Vector Space Model (VSM) - a vector, comprise of the keywords contained in the document, can represent a document. It is a powerful framework for examining and structuring the files. VSM model events can be divided into three types of stages: document indexing, the term weighting, and calculation of resemblance coefficients.

B. *Data Extraction*

Data extraction is the association of mechanically pulling out relevant information from large volume of texts. Extraction can acquire two forms; one is to identify the specific field of entity extracted such as name, date, or address, and the other one is to make out the parts of speech from text quantity by means of natural language processing (NLP) technology. Vantage Point applies NLP to parse text into the part(s) of speech. It utilized an amalgamation of semantic and syntactic analyses. It processes text inputs as follow:

C. *Data Pre-processing*

Data pre-processing, or data cleansing, is the algorithm that identify and erases the errors or inconsistency from the data and consolidates similar data in order to enhance the quality of consequent analysis. These cleaned details will then be feed to the examination process. Numerous methods could be utilized to clean the given data.

D. Data Analysis

Each document can be simply represented as a vector in a higher dimensional space. Therefore, dimensionality decrease methods are required to symbolize n-dimensional document data by a small number of significant dimensions.

E. Data Visualization

A common objective of analysis is to perceive meaningful underlying magnitude that permits the associate to describe observed similarities or dissimilarities (distances) in the midst of the investigated objects. This is completed by solving a minimization problem such that the distances among points in the conceptual low-dimensional space equivalent the given (dis)similarities as intimately as possible.

F. Applications of Text Mining

- Stopword - A stop word is a word that is distorted in the pre-processing of text. For instance words such as “is", “are", “you" and “me" can appear into view in any text document. The longer the certificate is the superior possibility of encountering them. These words have an effect on the model performance on the text mining task such as text categorization and text cluster. For the stop words removing process, the user must supply a “stop list"
- Stemming - The terms \procedure", “dispensation" and “process" originated from the root word “process". The query is, in a file, should we consider these words separately, or should we crumple them into a particular root form. The stemming process addresses this precise issue.

II. 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 while comparing with 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.

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 apply 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 the 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

Gouda K. et al (2010) presented a novel approach for mining frequent sequences, called Prism. It utilizes a vertical approach for enumeration and support counting, based on the novel notion of primal block encoding, which in turn is based on prime factorization theory. Via an extensive evaluation on both synthetic and real datasets, they show that Prism outperforms popular sequence mining methods like SPADE.

III. Problem Identification

There were some problems which have been identified from the previous works while researching about the topic. Those are explained in brief.

A. Intermediate Form

Intermediate forms with altering degrees of complexity are fit for different mining purposes. For a fine-grain domain-specific knowledge unearthing task, it is obligatory to perform semantic investigation to derive amply rich representation to detain the relationship between the objects or concepts described in the documents.

B. Multilingual Text Refining

Whereas data mining is principally language self-regulating, text mining involves an important language component. It is indispensable to develop text humanizing algorithms that procedure multilingual text documents and create languageindependent in-between forms. While for the most part text mining tools focus on dispensation English documents, mining from documents in former languages allows access to beforehand untapped information and offers a novel host of opportunities.

C. Domain Knowledge Integration

Domain knowledge, not catered for by any present text mining equipment, could play a significant role in text mining. Specially, domain information can be second-hand as near the beginning as in the manuscript refining stage. It is attractive to explore how one can obtain advantage of area information to progress parsing efficiency and obtain a more compact intermediate form. Domain knowledge possibly will also play a fraction in knowledge distillation.

D. Personalized Autonomous Mining

Current text mining merchandise and applications are still tools considered for trained knowledge specialists. Opportunity text mining tools, as part of the information management systems, should be willingly usable by technological users as well as organization executives. There have been some labors in developing systems that understand natural language queries and mechanically perform the suitable mining operations. Text mining tools could also appear in the form of clever personal assistants. Under the agent paradigm, an individual miner would learn a user's outline, conduct text mining operations mechanically, and forward information without requiring an plain request from the user.

IV. METHODOLOGY

The methodology which has been proposed for the solution of the problems identified in the project is as shown in the Figure 1.

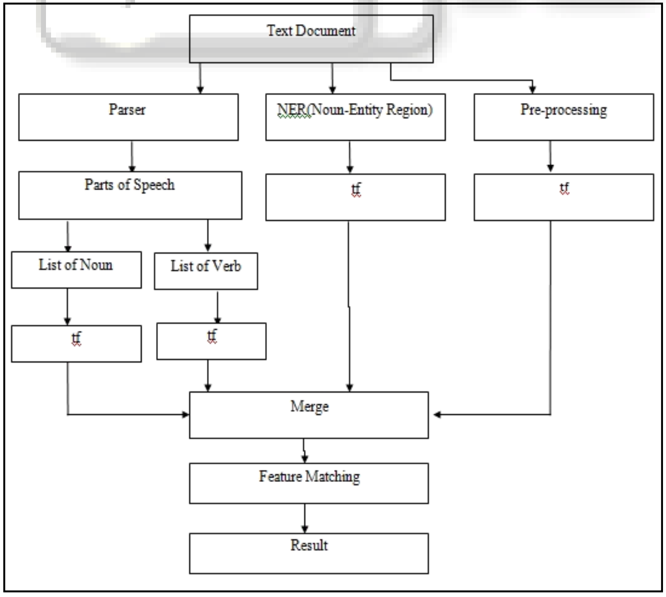


Fig. 2: Flowchart of the methodology

A. Step 1: Training the system to identify which class the text document belongs.

- The system is trained in three classes viz Historical Class, Constitutional Class and Geographical Class.
- The system has to be trained on the basis of the documents which are related to these predefined classes.
- The system is trained by extracting several keywords from the document related to a particular field then the keywords that are unique with respect to each other are stored to classify the given text document.
- Then after the text document is uploaded into GUI to find the class to which the text document belongs.

B. Step 2: Parse the text document using Stanford parser.

- Parsing or syntactic analysis is the process of analysing a string of symbols, either in natural language or in computer language, conforming to the rules of a formal grammar.
- The parsing is done by Stanford parser, which converts the text document into parse tree.
- One way to parse the text document is to parse each sentences of the text document individually and saving the output but the drawback is that it involves loading the Stanford parser several times.
- Therefore, the complete text document must be parsed at once to overcome the drawback.
- The headwords are saved for feature matching at later stage. Each word in the PTs is assigned with an ID to make system able of uniquely addressing words in the text document. This is required to avoid confusion amongst repeated words.
- The PT contains list of noun and verb which will be used in feature matching and by that concluding which class the given text document belongs.

C. Step 3: Apply Noun Entity Region (NER).

- Noun entity region (NER) (also known as entity identification, entity chunking and entity extraction) is a sub-task of information extraction that seeks to locate and classify elements in text into pre-defined category such as the names of persons, organizations, locations, expressions of times, quantities, monetary values, percentages, etc.
- To evaluate the quality of a NER system's output, several measures have been defined. While accuracy on the token level is one possibilities, it suffers from two problems: the vast majority of tokens in real-world text are not part of entity names as generally defined, so the baseline accuracy (always predict "not an entity") is extravagantly high, typically >90%; and mis predicting the full span of an entity name is not properly penalized.

D. Step 4: Pre – processing of the text document.

- Data may or may not have quality problems that need to be addressed before applying a data mining techniques.
- The Data may be irrelevant or duplicate, thus pre-processing is necessary.
- Pre-processing may be needed to make the text more suitable for text mining. There are a number of different tools and methods used for pre-processing.

- Text preprocesses is the initial step of text mining which reads one text document at time and process it. This step divides into following major three subtasks-

1) Tokenization

Generally text document contain multiples sentences. So this process divides whole sentence into words by removing comma, spaces, punctuations etc.

2) Stop Word Removing

This process removes stop words such as “the”, “are”, “a” or any tags like HTML tag etc.

3) Stemming

Stemming is applied after stop word removal by reducing the word to its root word. E.g. “playing”, “played” are stemmed to “play”.

- After pre-processing of the text document several headwords are generated. These headwords are saved to categorize the text document to the class it relates.

E. Step 5: Separate the list of noun and the list of verbs from the document.

- Separate the list of nouns and the list of verbs generated through that parser.
- There are lot of ‘ing’ and ‘es’ or‘s’ like words which have to be separated from the word to make it a proper word.
- There is a facility of word net , which recognizes the synonyms of the words and considers into the same word count, which is a very significant step in the method.
- Out of these, the nouns and the verbs are separated and made a parse tree.
- The list of noun is also generated by NER (Noun-Entity Region) which is termed as Main Part - I

F. Step 6: Merge all the features

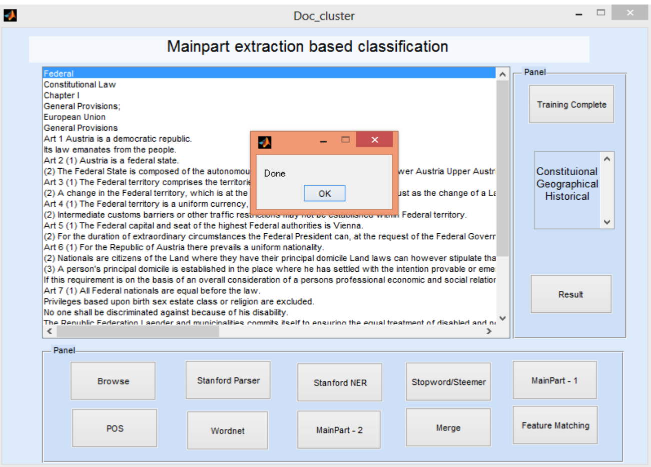
- After separating all the headwords viz noun and verb words generated by parser, noun words generated from NER (Noun-Entity Region) and the word generated after pre-processing, merge all the headwords extracted by system in order to identify that the given document belongs to which pre-defined class viz Historical Class, Constitutional Class and Geographical Class.

G. Step 7 – Match the features with the unique headwords of the predefined classes viz Historical Class, Constitutional Class and Geographical Class to find out the class to which the given document belongs.

- The Features are matched with the trained system, which matches the noun and verb present in the text document with that of the list of unique headwords of the classes which are pre-defined and tells us that in which class the text document belongs.
- If the text document doesn’t match with any of the classes, then it gives the result, no matching found.
- If the text document matches with any of the predefined classes, then it gives the name of the class as result to which the given text document belongs.

V. RESULTS

Based on the methodology discussed, we did got some extreme good and improved results while comparing with the previous works, which have been shown with the help of screenshots.

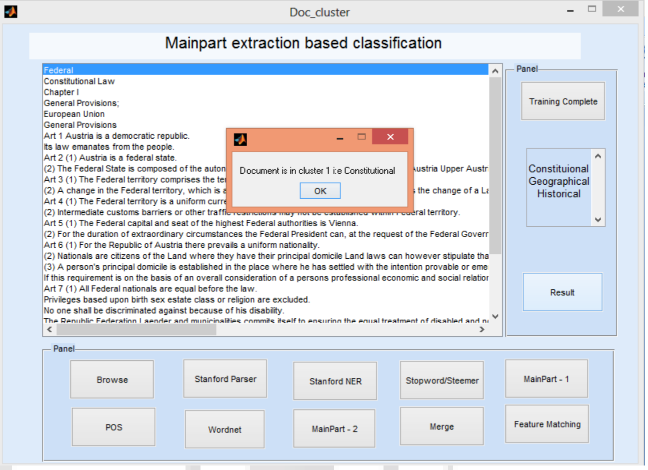


Federal
Constitutional Law
Chapter I
General Provisions;
European Union
General Provisions
Art 1 Austria is a democratic republic.
Its law emanates from the people.
Art 2 (1) Austria is a federal state.
(2) The Federal State is composed of the autonomou … wer Austria Upper Austr
Art 3 (1) The Federal territory comprises the territorie
(2) A change in the Federal territory, which is at the … ust as the change of a La
Art 4 (1) The Federal territory is a uniform currency,
(2) Intermediate customs barriers or other traffic res… Federal territory.
Art 5 (1) The Federal capital and seat of the highest Federal authorities is Vienna.
(2) For the duration of extraordinary circumstances the Federal President can, at the request of the Federal Goverr
Art 6 (1) For the Republic of Austria there prevails a uniform nationality.
(2) Nationals are citizens of the Land where they have their principal domicile Land laws can however stipulate tha
(3) A person's principal domicile is established in the place where he has settled with the intention provable or eme
If this requirement is on the basis of an overall consideration of a persons professional economic and social relatior
Art 7 (1) All Federal nationals are equal before the law.
Privileges based upon birth sex estate class or religion are excluded.
No one shall be discriminated against because of his disability.
The Republic Federation Laender and municipalities commits itself to ensuring the equal treatment of disabled and n

Done

OK

Fig. 3: Completion of Feature Matching as per Classification



Federal
Constitutional Law
Chapter I
General Provisions;
European Union
General Provisions
Art 1 Austria is a democratic republic.
Its law emanates from the people.
Art 2 (1) Austria is a federal state.
(2) The Federal State is composed of the auton … ustria Upper Austr
Art 3 (1) The Federal territory comprises the te
(2) A change in the Federal territory, which is a … the change of a La
Art 4 (1) The Federal territory is a uniform curre
(2) Intermediate customs barriers or other traffi … eral territory.
Art 5 (1) The Federal capital and seat of the highest Federal authorities is Vienna.
(2) For the duration of extraordinary circumstances the Federal President can, at the request of the Federal Goverr
Art 6 (1) For the Republic of Austria there prevails a uniform nationality.
(2) Nationals are citizens of the Land where they have their principal domicile Land laws can however stipulate tha
(3) A person's principal domicile is established in the place where he has settled with the intention provable or eme
If this requirement is on the basis of an overall consideration of a persons professional economic and social relatior
Art 7 (1) All Federal nationals are equal before the law.
Privileges based upon birth sex estate class or religion are excluded.
No one shall be discriminated against because of his disability.
The Republic Federation Laender and municipalities commits itself to ensuring the equal treatment of disabled and n

Document is in cluster 1 i:e Constitutional

OK

Fig. 4: Result showing that the text document belongs to Constitutional

VI. CONCLUSION AND FUTURE SCOPE

From the Research work, we came into a conclusion that mining the semantic information from free text document provides the enabling technology for a host to identify the class to which the text document belongs. The NER (Noun Entity Region) has been used to identify the noun keywords using classifier uniquely viz 3 – Class classifier, 4 – class classifier and 7 – class classifier. By using the concept of Parsing and NER, the text document has been classified to the predefined class to which the given text document belongs by merging the MP – I (List of noun) and MP – II (List of verb) and matched it with the stored headwords to conclude which class the document belongs to. The use of parser to convert the text document to parse tree increased the accuracy of the work. The Work is carried out with MATLAB as a Simulation tool, maybe there are some possibilities of increasing of accuracy rate if it is carried out by NS – 2 Simulator. As a future work, this solution can be enhanced by training the system in a broader way containing large number of documents and Classifiers. The System of Word net is also limited here, consisting of five synonyms of each word, which can be enhanced to observe more accuracy.

REFERENCES

[1] J.Han et al., Mining Frequent Patterns without Candidate Generation: A Frequent-Pattern Tree Approach. Data Mining and Knowledge Discovery, 8, 53–87, 2004. Kluwer Academic Publishers. Netherlands.

[2] M.Hu, and B.Liu, Mining and Summarizing Customer Reviews, in: Proceedings of ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD’04), USA, 2004, pp. 168 – 177.

[3] B.Pang, L.Lee, and S.Vaithyanathan, Thumbs up? Sentiment Classification Using Machine Learning Techniques, in: Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing (EMNLP’02), USA, 2002, pp.79 –86.

[4] B.Liu, M.Hu and J. Cheng, Opinion Observer - Analyzing and comparing opinions on the Web, in: Proceedings of the 14th International Conference on World Wide Web (WWW’05), Japan, 2005, pp. 342-351.

[5] A.M.Popescu and O. Etzioni, Extracting Product Features and Opinions from Reviews, Proceedings of the 2005 Conference on Empirical Methods in Natural Language Processing (EMNLP’05), Canada, 2005, pp. 339 – 346.

[6] X.Ding, B. Liu and S.Y.Philip, A Holistic Lexicon-Based Approach to Opinion Mining, in: Proceedings of the first ACM International Conference on Web search and Data Mining (WSDM’08), California, USA, 2008, pp. 231-240.

[7] M. Abulaish, Jahiruddin, M. N. Doja and T. Ahmad, “Feature and Opinion Mining for Customer Review Summarization”, PReMI 2009, Lecture Notes in Computer Science, vol. 5909, pp. 219–224, Springer-Verlag Berlin Heidelberg 2009. [10] M. Atzori and C. Zaniolo. Swipe: searching wikipedia by example. In WWW, pages 309–312, 2012.

[8] B.Pang, and L. Lee, A Sentimental Education: Sentiment Analysis Using Subjectivity Summarization Based on Minimum Cuts, in: Proceedings of ACL 2004, 2004, pp. 271-278.

[9] R. Alur and P. Madhusudan. Adding nesting structure to words. In Developments in Language Theory, 2006.

[10] M. Atzori and C. Zaniolo. Swipe: searching wikipedia by example. In WWW, pages 309–312, 2012.

[11] E. Charniak and M. Elsner. Em works for pronoun anaphora resolution. In EACL, pages 148–156, 2009.

[12] S. A. Cook. The complexity of theorem-proving procedures. In STOC, pages 151–158, 1971.

[13] H. Cunningham, D. Maynard, K. Bontcheva, and V. Tablan. Gate: an architecture for development of robust hlt applications. In Recent Advanced in Language Processing, pages 168–175, 2002.

[14] B. J. Grosz, S. Weinstein, and A. K. Joshi. Centering: a framework for modeling the local coherence of discourse. Comput. Linguist., 21(2):203–225, June 1995.

IJSRD