Efficient Retrieval of Text for Law Reports using Enhanced Data Mining Algorithm

KEYWORDS Data Mining. Text mining

Siddharth Arora
Department of Computer Science & Engineering
Ambala College of Engineering and Applied Research,
Devasthali, Ambala Cantt-133001, Haryana, INDIA

Parneet Kaur
Department of Computer Science & Engineering
Ambala College of Engineering and Applied Research,
Devasthali, Ambala Cantt-133001, Haryana, INDIA

ABSTRACT *Data mining, a branch of computer science [1], is the process of extracting patterns from large data sets by combining methods from statistics and artificial intelligence with database management. Data mining is seen as an increasingly important tool by modern business to transform data into business intelligence giving an informational advantage.. Recent methods of soft clustering can exploit predictive relationships in textual data. This paper presents a technique for using clustering data mining algorithms to increase the accuracy of the scattered code. In this proposed work, the work is done to enhance the scattering pattern of large java data sets.*

I. INTRODUCTION

Data mining is the process of sorting through large amounts of data and picking out relevant information. It is usually used by business intelligence organizations, and financial analysts, but is increasingly being used in the sciences to extract information from the enormous data sets generated by modern experimental and observational methods. It has been described as “the nontrivial extraction of implicit, previously unknown, and potentially useful information from data” and “the science of extracting useful information from large data sets or databases” [3]. Text mining, sometimes alternately referred to as text data mining, refers generally to the process of deriving high quality information from text. High quality information is typically derived through the division of patterns and trends through means such as statistical pattern learning. Text mining usually involves the process of structuring the input text (usually parsing, along with the addition of some derived linguistic features and the removal of others, and subsequent insertion into a database), deriving patterns within the structured data, and finally evaluation and interpretation of the output. ‘High quality’ in text mining usually refers to some combination of relevance, novelty, and interestingness. Typical text mining tasks include text categorization, text clustering, and concept/entity extraction, production of granular taxonomies,sentiment analysis, document summarization, and entity relation modeling

II. METHOD

The input to the proposed system is a collection of law reports. Law reports consist of two sections; namely the head and the detail section. The head section summarizes the whole law report and the detail section contains the detailed information about the case. Only the head section is used for automated processing as it contains sufficient details for the purpose. The proposed system consists of two main components, namely;

a. The mining process
b. The research process

The mining process is the main process in the framework and to be completed prior to the research process. The mining process is carried out on the entire collection of the law reports of the repository. In this process, each document is analyzed and information that should be used for legal research is recorded in the processed law reports repository. Then the research process is carried out on the processed law reports. In this process, the text block is analyzed and the required information is extracted and compared with each law report to identify the matching reports.

The proposed approach uses text mining. More precisely, it uses terms level text mining, which is based on extracting meaningful terms from documents [11]. Each law report is represented by a set of terms characterizing the document.

III. COMMON TECHNIQUES OF DATA MINING

There are many techniques of data mining. The most common techniques used in the field of data mining are followings.

1) Artificial neural networks

Non-linear predictive models that learn through training and resemble biological neural networks in structure. This predictive model uses neural networks and finds the patterns from large databases.

2) Decision trees

Set of decisions are represented by Tree-shaped structures. These decisions generate rules for the classification of a dataset under the large databases. Specific decision tree methods include Classification and Regression Trees (CART) and Chi Square Automatic Interaction Detection (CHAID).

3) Genetic algorithms

Optimization techniques that use processes such as genetic combination, mutation, and natural selection in a design based on the concepts of evolution.

4) Nearest neighbor method

A technique that classifies each record in a dataset based on a combination of the classes of the k record(s) most similar to it in a historical dataset (where k ³ 1). This is sometimes called the k-nearest neighbor technique.

5) Rule induction

The extraction of useful if-then rules from data based on statistical significance between different records of database.

Many of these technologies have been in use for more than a decade in specialized analysis tools that work with relatively small volumes of data. These capabilities are now evolving to integrate directly with industry-standard data warehouse and OLAP platforms [8]. The appendix to this white paper provides a glossary of data.

6) Fuzzy C Means

Here algorithm is responsible for extracting keywords present in the full text biomedical article store these keywords in a relation. Then the actual work of algorithm begins, it starts clustering of keywords. The algorithm initially picks some keywords that are extracted . It groups the full text articles based on these keywords. It means each cluster contains only those articles which contain that keyword as their part. Then it starts using fuzzy C mean clustering to combine the clusters together on some similarity measure. Here we combine two clusters if their similarity measure is greater than or equal to a specified threshold value. The proposed Algorithm repeats this process until no more changes are made to the clusters. Finally the proposed algorithm stores all the clusters in an xml file. Here our motive to extract all the full text articles which may be relevant for the user providing the search string, for this out of all clusters the cluster with largest number of articles is our target. the enhanced part of the algorithm is integrated with dictionary words as well in case the user entered the wrong word the enhanced model of algorithm will compare the words present in dictionary and find the exact word of the related keyword and displays the correct result as previous one.

7) Sequential Pattern Mining

Sequential pattern mining has become an essential task with broad applications. Most sequential pattern mining algorithms use a minimum support threshold to prune the combinatorial search space. This strategy provides basic pruning; however, it cannot mine correlated sequential patterns with similar support and/or weight levels. If the minimum support is low, many spurious patterns having items with different support levels are found; if the minimum support is high, meaningful sequential patterns with low support levels may be missed. We present a newalgorithm, weighted interesting sequential (WIS) patternmining based on a pattern growth method in which new measures, sequential s-confidence and w-confidence, aresuggested. Using these measures, weighted interesting sequential patterns with similar levels of support and/or weight are mined. The WIS algorithm gives a balance between the measures of support and weight, and considers correlation between items within sequential patterns. A performance analysis shows that WIS is efficient and scalable in weighted sequential pattern mining.

IV. METHODOLOGY

Algorithm

1. Read the next article in the list of Law cases
2. Read the full text article
3 Extract the keywords from the article using weighted algorithm.
4. Refer to the Law report and discard the irrelevant keywords
5. Put the data in following relation so that the full text can be retrieved later using keywords only
6. Go to step 1 and repeat till all the articles in the list of biomedical articles are processed.
7. Use the fuzzy c-means algorithm to create clusters on keywords
8. Save the article clusters in form of an XML file(containing articles IDs).

V. CONCLUSION

This paper presents the results of the research carried out to develop a framework to automate the often tedious time consuming process of legal research. The end result of the research is a framework which is based on a combination of several text mining techniques. Finally the framework developed was tested for accuracy using a prototype application and fundamental rights case records. Accuracy of the results in terms of precision and recall were shown to be very high. As in our work, this can be extended to handle all types of law reports in addition to the fundamental rights cases. The accuracy can be further enhanced by using a comprehensively updated stop words list and a set of predefined terms to be introduced along with the dropping of candidate terms. The weighting scheme can also be upgraded to include context information.

VI. ACKNOWLEDGMENTS

Our thanks to the experts who have contributed towards development of the thesis.

REFERENCE [1] Efficient Retrieval of Text for Biomedical Domain using Data Mining Algorithm. (IJACSA) International Journal of Advanced Computer Science and Applications, Vol. 2, No. 4, 2011 | [1] Clifton, Christopher (2010). "Encyclopedia Britannica: Definition of Data Mining". Retrieved 2010-12-09. | [2] Han, J., & Kamber, M., Data Mining Concepts and Techniques. CA Morgan Kaufmann, 2001. | [3] Badgett RG: How to search for and evaluate medical evidence. Seminars in Medical Practice 1999, 2:8-14, 28. | [4] Richardson J: Building CAM databases: the challenges ahead. J Altern Complement Med 2002, 8:7-8. | [5] Kantardzic, Mehmed (2003). Data Mining: Concepts, Models, Methods, and Algorithms. John Wiley & Sons. ISBN 0471228524. OCLC 50055336 | [6] Miller, H. and Han, J., (eds.), 2001, Geographic Data Mining and Knowledge Discovery, (London: Taylor & Francis). | [7] Manu Aery, Naveen Ramamurthy, and Y. Alp Aslandogan. Topic identification of textual data. Technical report, The University of Texas at Arlington, 2003. | [8] Pavel Berkhin. Survey of clustering data mining techniques. Technical report, Accrue Software, San Jose, CA, 2002. | [9] Cecil Chua, Roger H.L. Chiang, and Ee-Peng Lim. An integrated data mining system to automate discovery of measures of association. In Proceedings of the 33rd Hawaii International Conference on System Sciences, 2000. | [10] George Forman. An extensive empirical study of feature selection metrics for text classification. J. Mach. Learn. Res., 3:1289-1305, 2003. | [11] Rayid Ghani. Combining labeled and unlabeled data for text classification with a large number of categories. In IEEE Conference on Data Mining, 2001