Knowledge Discovery from Legal Documents Dataset using Text Mining Techniques

Rupali Sunil Wagh
Christ university, Hosur Road,
Bangalore, 560029

ABSTRACT
Last few decades have witnessed exponential increase in the use of IT which has resulted into large amount of data being generated, stored and searched. Data may be highly structured stored as records of a DBMS, or may be totally unstructured like blog posts or plain text documents. With the abundance of information being available as text documents, the issue of retrieval of knowledge from such unstructured dataset is posing new challenges to the research community. Legal document analysis is one domain which generates and uses text information in semi structured as well as unstructured form. The process of legal reasoning and decision making is heavily dependent on information stored in text documents. Text Mining (TM) is defined as the process of extracting useful information from text data. Legal text documents are stored using natural languages. For efficient analysis of such documents, text mining, a specialized branch of machine learning can be suitably used. Text mining – which "mines text", is heavily associated with natural language processing and Information Retrieval. TM techniques can be used for extracting relevant knowledge from stored legal documents. The extracted knowledge is used to simplify the preparation of case base, facilitate in decision making and legal reasoning or for automatic identification of legal arguments. Research in the fields of information extraction, natural language processing, artificial intelligence and expert system has augmented text mining process for enhancing the knowledge discovery process in this domain. This paper proposes a study which is aimed at grouping of legal documents based on the contents without taking any external input using unsupervised text mining techniques.

General Terms
Legal research, legal domain, keyword based search.

Keywords
Key words – Text mining, legal document search, legal databases, unsupervised learning.

1. INTRODUCTION
Text mining is referred to as an extension to traditional data mining to handle unstructured text data. Substantial amount of data today is stored in text databases and not in structured databases. This very fact makes text mining research increasingly important. As pointed out by Kong [1], the huge set of words and varied rules of sentence construction in natural language along with uncertainty and ambiguity in the text makes TM a challenging task. Kong[1] has also mentioned the language oriented aspect of such analysis which makes it very different and complex as compared with the analysis of structured data.

Process of legal reasoning and argumentation is based on information extracted from a variety of documents. TM research in this domain aims at facilitating legal logic development by providing deeper and "intelligent" insights into the available data. Text clustering can group the documents based on the contents of documents without taking any input from the user. Such grouping can be very useful to filter task irrelevant data and thereby improving the search operation in legal study.

.

2. RELATED WORK
Text or documents is very common and important source of information, often semi structured or unstructured. [2]points out the similarities and differences between techniques popularly used for text analysis, namely information retrieval, natural language processing, document classification and clustering along with the comparative study of document clustering algorithms and mention about TM tools.

Legal information is a huge collection of various documents as mentioned in the introduction to study. Legal search and legal reasoning are two major processes of legal domain. Applications of DM in this domain were proposed years back. Most of these applications are aimed at improvement of the legal document search process. Application of self-organizing maps for legal document clustering was proposed in [3] back in 1997. With the increased access and availability of the data, the applications TM techniques in legal domain have gained more popularity in last decade and primarily used to improve the search result of which is the backbone for legal reasoning. [4] Has suggested a system based on information extraction techniques for retrieving information form legal text documents written in different styles of writing formatting and footnoting. IBM researchers [5] have proposed E-discovery reference model which uses TM and information retrieval methods and augmenting it with semantic and syntactic analytics techniques to improve the efficiency of knowledge discovery from legal document set. A novel approach of transforming legal documents to XML documents was proposed in [6]. Though applicable to only limited countries, this approach aims at using unstructured text for generation of metadata templates for legal search that could be used further processing of the text on the net. Legal document segmentation for improving the search result accuracy and overall search task complexity is proposed in [7], where the inherent informal structure of such documents is used for segmentation of documents. [8] Has proposed a very interesting approach of automatically creating an expert-witness database by analyzing text. [9] Discusses about a comparative analysis of man Vs computer document review in legal domain and provides further insights into the challenges and scope for further improvement. Insufficiencies of traditional TM techniques are analyzed in [10].

The literature thus emphasizes on the suitability and the aptness of text mining techniques for facilitating the processes for knowledge workers of legal domain. It is also evident that improving the accuracy and efficiency of the text search in this domain is challenged by the enormity, inherent complexity and context sensitivity of the data. Sound background knowledge is the major prerequisite for the keyword based search that is popularly used in this domain.

With newer TM techniques and models, researchers are trying to minimize user's input and facilitate more intelligent and automated search of legal text documents. Proposed study aims at grouping of legal text documents using unsupervised learning technique. Such groupings could be useful in filtering irrelevant documents and automatic identification of sub categories of a concept, thereby enhancing the search process for legal domain analysis.

3. DOMAIN DETAILS

3.1 Legal document search

A Legal professional needs to perform extensive search through various legal documents for every case in order to arrive at a conclusion. These documents include judgments, notifications, acts treaties etc. The documents could further be subdivided into different categories based on the judicial system of the country. (For example high court, Supreme Court, cyber law, consumer law, corporate law etc).

With electronic availability of legal documents in legal document databases, this legal research process today is totally computer assisted process. Many online legal databases like lexisnexis, manupatrawhich store and provide the information related to legal domain. All information related to legal domain is available in these databases as semi structured documents and can be searched primarily using a keyword based search. With the advances in information extraction and information retrieval technology, the search process is getting improvised. But, because of the volume and the inherent complexity of text data, finding relevant and required document from the legal database without proper background knowledge is still a difficult task.

3.2 Text mining

We all tend to respond in natural languages. Since it does not need any training, free text is the easiest way to input data. Thus data as unstructured or semi structured text is available in plenty. But ironically, this most common form of data is the most complex to analyze. Along with the unstructuredness, these complexities are due to the semantic intricacies, ambiguities and context sensitivity inherent to natural languages. These issues complicate the process of knowledge discovery from text KDT. TM – aims at knowledge discovery from textual dataset. Numerous algorithms and techniques have been proposed to improve the process of knowledge extraction from text data.

While, basic TM activities of text classification, clustering and association mining still find their relevance in KDT process, researchers have proposed combination of such approaches with other techniques for improvisation of processes. Novel approach of using genetic algorithm and semantics for automatic generation of training and testing dataset from given input has been proposed in [11]. Ontology is also emerged as a very important concept that along with TM techniques could be used for KDT effectively [12,13].

The process of knowledge extraction from text data takes inputs form various related fields. Information retrieval and information extraction is at the core of the TM activities. Language oriented nature of this process requires natural language processing techniques to handle the semantics and the ambiguities present in the text information. Expert system and artificial intelligence help in building the knowledge base to automate the process. Each of these fields individually or in combination, with its rich set of techniques, contributes substantially in this process of knowledge extraction form text databases.

As pointed out in earlier sections, easy availability of text data is posing new challenges in the field of Knowledge discovery from stored data. With the varied set of algorithms and techniques, TM, since it's inception has been catering to knowledge discovery in many domains. Undoubtedly, it will continue doing so with constantly evolving tools and techniques. Research in other related fields of NLP, IR and artificial intelligence has always complemented TM to enhance the KDT process.TM has emerged as an important research field because of its relevance in today's computational dynamics.

4. METHODOLOGY

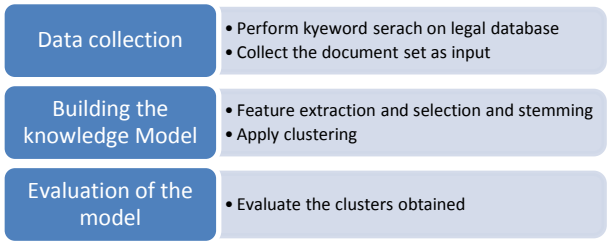

Figure 1 – Legal document grouping using clustering

As shown in the figure, the proposed study would be carried out in following steps-

1. Data collection
   The legal document dataset can be collected from legal databases. This dataset would actually be result of keyword search based on particular concept. For example – keyword based search for the concept education selects many documents as result. These documents can be considered as input dataset.
2. Building knowledge model using unsupervised mining technique – clustering
   This step follows traditional text mining activities - 1) Preprocessing deals with feature extraction, feature recognition and stemming etc. 2) Grouping of the documents using algorithm well suited to the data. Though neural network and machine learning algorithms are widely used approaches for text clustering, actual choice of the algorithm will be decided in due course of study. For example the input dataset as described in step 1 might have documents that could be grouped further like higher education, primary education etc. Clustering is aimed at performing this grouping based on the contents of the documents without any external input which will lead into automatic identification of subtopics of a concept.
3. Evaluation of knowledge model
   The developed model would be evaluated for its performance.

5. CONCLUSION

Data mining has been contributing to numerous domains and problems for knowledge insights into the data. Text Mining, the extension to DM has gained popularity in last decade because of the availability of large volume of text data. Information stored in text is of great commercial value. TM research aims at improving human's ability to comprehend huge data to provide better insights by analyzing a document or documents.

There are many domains that are primarily based on the information stored as text and legal domain being one of

them. For a legal analyst, every case at hand is a research problem. Legal or judicial reasoning is based on exhaustive search which aims at argument invention. The volume and the complexity of the documents to be searched and analyzed, makes the above mentioned task very cumbersome. Most of the search options offered today are keyword based. Researchers have proposed TM method and techniques for simplifying this process. The study proposes application of unsupervised text mining technique –clustering for grouping documents to enhance the document search.

6. REFERENCES

[1] Kong Yanqing and Guoliang Shi Guoliang, Advances in Theories and Applications of Text Mining. The 1st International Conference on Information Science and Engineering (ICISE2009)

[2] K. A Vidhya and Aghila G, "Text Mining Process, Techniques and Tools : an Overview*",* International Journal of Information Technology and Knowledge Management,July-December 2010, Volume 2, No. 2, pp. 613-622

[3] Merkl Dieter and Schweighofer Erich *"*En Route to Data Mining in Legal Text Corpora: Clustering, Neural Computation, and International Treaties*",* 0-8186-8147-0/97 IEEE 1997

[4] Cheng Tin Tin, Leonard Cua Jeffrey, Davies Tan Mark, Gerard Yao Kenneth and EditaRoxas Rachel. Information Extraction from Legal Documents**,** 2009 Eighth International Symposium on Natural Language Processing, 2009 IEEE

[5] Joshi Sachindra, Deshpande, Prasad M and Hampp Thomas. Improving the Efficiency of Legal E-Discovery. 2011 Annual SRII Global Conference, DOI 10.1109/SRII.2011.97

[6] Ismael Hasan, Javier Parapar, Roi Blanco.Segmentation of legislative documents using a domain-specific lexicon. 19th International Conference on Database and Expert Systems Application, DOI 10.1109/DEXA.2008.45

[7] Palmirani Monica and BrighiRaffaella. Metadata for the Legal Domain**.**Proceedings of the 14th International Workshop on Database and Expert Systems Applications (DEXA'03)

[8] Dozier Christopher and Jackson Peter, "Mining text for expert witnesses",2005 IEEE

[9] Roitblat Herbert L., Kershaw Anne and Oot Patrick. "Document Categorization in Legal Electronic Discovery: Computer Classification vs. Manual Review",JOURNAL OF THE AMERICAN SOCIETY FOR INFORMATION SCIENCE AND TECHNOLOGY, 61(1):1–11, 2010

[10] S´anchez D., Mart´ın-Bautista M.J., Blanco I., C. Justicia de la Torre. Text Knowledge Mining: An Alternative to Text Data Mining.2008 IEEE International Conference on Data Mining Workshops

[11] John Atkinson-Abutridy, Chris Mellish, and Stuart Aitken, "Combining Information Extraction With Genetic Algorithm for Text Mining", IEEE INTELLIGENT SYSTEMS

[12] Li Yaxiong, Zhang Jianqiang, Dan Hu. Text Clustering Based on Domain Ontology and Latent Semantic Analysis.2010 International Conference on Asian Language Processing, DOI 10.1109/IALP.2010.55

[13] MircoSperetta and Susan Gauch. Using Text Mining to Enrich the Vocabulary of Domain Ontologies. 2008 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology, DOI 10.1109/WIIAT.2008.288