International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 3 Issue 2, February - 2014

Exploratory Analysis of Legal Documents using Unsupervised Text Mining Techniques

Rupali Sunil Wagh
Christ University, Bangalore

***Abstract*--Profuse availability of digital data has posed a challenging problem of information overload in all domains. Information processing in legal domain has also undergone fundamental changes over last few decades. Two main issues concerning information in this domain are – a) refining the techniques and methods for handling complexity of knowledge in this domain b) Suitable ways to store and retrieve the information. The huge amount of information requires more intelligent machines. The paper proposes a novel approach for grouping case notes(abstracts) of legal document dataset for better grouping**

Keywords— Legal documents, clustering, document grouping, topic identification Introduction (*HEADING 1*)

I. INTRODUCTION

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 domain generates huge information in the form of text and documents. Legal information can be categorized under various headings like court transcripts, verdicts, statements and affidavits etc. Such documents are repositories of useful information regarding the interpretations of law and a legal researcher has to study such documents for preparing a case base. Many online legal databases provide easy access to such legal documents. Though most of these searches are keyword searches and follow Boolean retrieval model, simple options facilitating comprehensive search are available.

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. TM research in this domain aims at facilitating legal logic development by providing deeper and "intelligent" insights into the available data. The study proposes application of unsupervised text mining technique –clustering for grouping documents to enhance the document search. 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

II. RELATED WORK

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. Text or documents is very common and important source of information, often semi structured or unstructured. 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 are elaborated in [2].

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]. Legal documents generally belong to multiple categories. The information stored in legal documents is very different and is carefully written by experts. Considering these characteristics of the domain soft clustering techniques are proposed in [11]. SimRank algorithm [12], an approach based on relationship between law entities in legal text is also proposed.

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.

III DATA DESCRIPTION

Manupatra, one of the leading online legal database in India. It has user friendly interface and provides many search options like legal search, Bench search, manu search etc. Depending on the category of the document, the back end database records various field values thereby transforming the unstructured document into structured information. A user can query the database by providing appropriate keyword. Phrase queries can also be handled by the retrieval system supporting more flexible search. The precision and recall of these search options are sensitive to framing of search query.

The data required for the study are the documents regarding judgment retrieved after giving a search query. The documents are divided into following parts-

- Catchwords
- Date of the judgment
- Details about the court and the bench
- Appellants
- Respondent
- Judges
- Subject (categorization viz civil)
- Rules/Order
- Cases Referred
- Disposition
- Case Notes (Abstract of the case)
- Detailed judgment given by the court

Each of the documents is identified using a set of 15 to 25 catchwords. In the keyword based retrieval system, huge number keywords affect the recall of the system. The study has considered only the catchwords and the case notes for cluster analysis. Subdivisions may vary based on the judicial system of the country. (For example high court, Supreme Court, cyber law, consumer law, corporate law etc).

IV EXPERIMENTAL SETUP

Clustering is the most common form of unsupervised learning. It requires no human expert to group the data which means that the grouping is performed on the basis of inherent similarities and differences among the documents. Since the documents belonging to one cluster are "similar" to each other, text clustering can be applied aptly in information retrieval

A. Document clustering for topic identification

As highlighted in the beginning of the document, text clustering is the most commonly used text mining technique because of its unsupervised nature. With its wise range of algorithms, this unsupervised and most natural way of grouping objects has been very predominant in evry domain. "Dissimilarity" is the basic concept around which the clustering process works and with varies measures of dissimilarity like Euclidian, cosine, Manhattan etc it caters to various domain needs [13].

While grouping objects is the very basic step in any analysis, identifying the topics of the grouped data is equally important. Document clustering can be effectively used for identification of relevant categories of the clusters[14,15]. It is also highlighted that cosine measures are the most widely used measures for document clustering problem.

B. Methodology

47 Documents retrieved form the online legal database using the query "patents act" are downloaded. Every retrieved document is then divided into two parts – Catchwords and Case Notes The processing of these subparts would be done independently for analyzing and grouping. Figure 1 shows the steps of the methodology being followed for the study

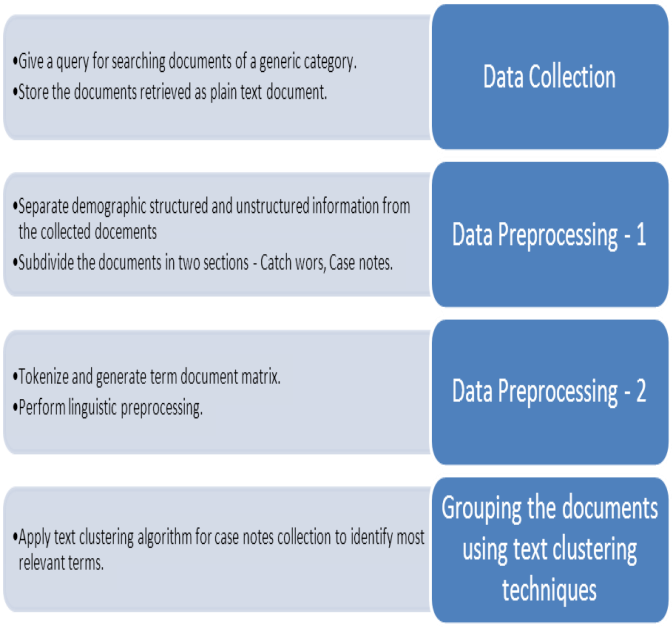

Figure 1

Linguistic Preprocessing, the step 3 of the shown figure forms the basis for all the TM functionalities. In its basic for linguistic preprocessing includes-

1.Case Conversion
2. Stopword Removal
3. Punctuation marks Removal
4. Stemming
5. Index Creation

Once the index is created text clustering is applied. There are many approaches like N gram index, byword index, phrase index The study has used single words tokens.

Cluster analysis is applied to the collection of case notes

which is written in plain English language. Algorithm spherical kmeans, the variation of kmeans algorithm to include cosine measures is used for the cluster analysis

RESULTS AND DISCUSSION

The results obtained by applying clustering algorithm skmeans are summarized in the following tables.

Table1 – skmeans clusters with k=3
Number of clusters 3, Cluster Size 21,12,14

Cluster No	Most Prominent Terms in The cluster with the frequency			
Cluster 1	**Patent(480)**	**Invention(230)**	**Application(208)**	**Claim(177)**
Cluster 2	**Mark (86)**	**Trade(74)**	**Copyright(66)**	**Infringement(65)**
Cluster 3	**Suit (53)**	**File (39)**	**Application (34)**	**Appeal (31)**

Table 2 skmeans clusters with k=4
Number of clusters 4, Cluster Size 23,9,9,6

Cluster No	Most Prominent Terms in The cluster with the frequency			
Cluster 1	**Patent(480)**	**Invention(230)**	**Application (208)**	**Claim (177)**
Cluster 2	**Mark (96)**	**Trade(78)**	**Copyright(64)**	**Infringement(61)**
Cluster 3	**Appeal (28)**	**Jurisdiction (28)**	**Cause (24)**	**Suit (24)**
Cluster 4	**Design(63)**	**Register(26)**	**Case(25)**	**Cancel(20)**

As it can be seen from the tables, the clusters reflect the relevant terms for the groups. This precise information regarding the document groups helps in specific search and getting the most relevant documents.

VI. CONCLUSION AND FUTURE WORK

This paper has presented document cluster analysis as a tool to enhance search in legal domain. Plain text documents which contain abstracts of cases can be further subdivided into different groups to identify more specific and relevant based on the abstracts than the system provided keywords.

This approach is based on the cosine similarity model and does not take into account any domain specific concepts. If the basic clustering is enhanced with domain ontology, underlying concepts and entities could be considered for grouping the documents. Such cluster analysis can be more comprehensive and will result in better results. Application of sematic concepts and related feature reduction technique may also improve the quality of obtained clusters.

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 Edita Roxas Rachel, "*Information Extraction from Legal Documents***",** 2009 Eighth International Symposium on Natural Language Processing, 2009 IEEE

[5] Joshi Sachindra, DeshpandePrasad 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 Brighi Raffaella, "*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] Qiang Lu, William Keenan, Jack G. Conrad, Khalid Al-Kofahi, "*Legal Document Clustering with built in Topic Segmentation*", CIKM'11, October 24–28, ACM 978-1-4503-0717

[12] Biao Fan, Tao Liu, He Hu and Xiaoyong , "*Law Text Clustering based on Referential Relations",* 2010 IEEE, 978-0-7695-4106-8/10, DOI 10.1109/ChinaGrid.2010.22

[13] Rui Xu, Donald Wunsch, "*Survey of Clustering Algorithms*", IEEE TRANSACTIONS ON NEURAL NETWORKS, VOL. 16, NO. 3, MAY 2005

[14] Antoine Naud and Shiro Usui , "*Exploration of a collection of documents in neuroscience and extraction of Topics by Clustering*" , IDEAL'07 Proceedings of the 8th international conference on Intelligent data engineering and automated learning

[15] Chris Clifton, Robert Cooley, Jason Rennie "*TopCat: Data Mining for Topic Identification in a Text Corpus*" , IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 16, NO. 8, AUGUST 2004