DATA MINING AND KNOWLEDGE REPRESENTATION IN DOCUMENT PROCESSING FOR LAW FIRMS-A Survey

Ankita¹, Dhanashree², Diksha³, Mangesh⁴, Pranoti⁵, Reena Thakur⁶
Guru Nanak Institute of Engineering and Technology, Nagpur
¹ankitab71@yahoo.in ²dhanashreekakde@gmail.com ³patel.diksha02@gmail.com
⁴mangeshkshirsagar7@gmail.com ⁵tanu.mane@gmail.com ⁶rina151174@gmail.com

ABSTRACT
Streamlining document process and increasing efficiency are fundamental concerns for any organization. All the official documents have been processed manually but it seems very difficult if someone needs to have particular information from particular document. Data mining in document processing using various techniques, such as classification, clustering has been developed to handle the unstructured documents. This paper gives quick summary of some of the most important developments in data mining in document processing research. In this paper we have presented a survey which covers techniques and methods that promise to enable us to mine documents information from vast amount of databases. Our aim was mainly on machine learning techniques on the basis of their usage and importance for document mining.

Keywords: data mining, clustering, classification, machine learning, k-means, fuzzy k-means.

1. INTRODUCTION
Law firms dealt with countless amounts of various paper documents every day as huge mass of data is being accumulated in the data repository. The costs associated with handling and processing these documents is exorbitant. Generally there is a huge gap from the stored data to the knowledge that could be constructed from the data. This transition i.e. document management won’t occur mechanically, that’s where document process in Data Mining comes into picture. If we use Manual data analysis, it creates a bottleneck for large data analysis. A number of Data Mining techniques are developed to mine the vast amount of data. Document databases are rapidly growing due to the increasing amount of information available in electronic forms, like, e-mail, CD-ROMs, and other World Wide Web.

Document Mining combines many of the techniques of information extraction such as information retrieval, and natural language processing and document summarization with the methods of data mining [04]. The key use for document mining is to extract previously unknown knowledge locked away in a bulk of text [02].

All the official documents have been processed manually but it seems very difficult if someone needs to have particular information from particular document. Data mining in document processing using various techniques, such as classification, clustering has been developed to handle the unstructured documents. But, traditional document processing techniques become insufficient for the increasingly huge amount of text data. However, only a small fraction of the available documents will be similar to a given individual. For analyzing and extracting useful information from the data, it is hard to formulate effective queries, without actually knowing what could be in the documents. Thus, document processing in Data mining has become an increasingly popular and essential theme in Data Mining.

As we go down, section II describes the literature review, section III and IV describes problem definition and issues, section V describes proposed methodology and finally section VI describes implementation model and references.

2. LITERATURE REVIEW
This paper discusses advantages of document classification methods for organizing explicit knowledge. The objective of document classification is to reduce the detail and diversity of data and the resulting information overload by grouping similar documents together. The term “document classification” is often used to consider two types of analyses: document categorization and document clustering. The distinction is that categorization is a form of supervised and clustering an unsupervised approach of grouping textual objects.

In [1] classification of text document has been studied which is a supervised technique that uses labeled training data to learn the classification system and then automatically classifies the remaining text using the learned system. In this paper, the author proposes a mining model consists of document-based concept analysis. Feature selection is performed after extracting feature vector for each new document. It is then followed by K-Nearest Neighbor classification.

Document clustering refers to unsupervised classification of documents into groups in such a way that the documents in a cluster are relevant, whereas documents in different clusters are irrelevant [7]. In this paper application of K-means, heuristic K-means and fuzzy C-means algorithms for clustering text documents has been introduced. Different representations (tf, tf.idf & Boolean) and different feature selection schemes (with or without stop word removal & with or without stemming) are examined.

Document processing is an interdisciplinary field which involves Information Extraction, Categorization, Information Retrieval, Clustering, Text Understanding, Concept Linkage, Database Technology, Topic Tracking, Machine Learning, and Data Mining [9]. Document Mining tools/applications intend to capture relationship between the one document with the other document. They can be roughly organized into two groups. One group focuses on document functioning to organize documents based on their priority. The other group focuses on document analysis functions to analyze the content of the documents and discover relationships between concepts or entities described in the documents.

3. PROBLEM DEFINITION
In a manual system, the user must have the prior knowledge of where an existing document is stored and what its name is. If the document belongs to the original owner then it is fare enough but if it has prepared by someone else then it will take sufficient amount of time.

The manual system will propagate to human errors. A user may have stored a particular document in the wrong place or may have forgotten the name of the document or even placed an entire directory to some new location without even being aware of it. In such situation if some other person rather than the original author tries to search for the document, then difficulties are compounded. A user may have to search in four or five places before finding a document or even sometimes unable to find it at all.

4. ISSUES IN DOCUMENT PROCESSING
Existing document processing techniques become inadequate to handle large databases containing high volume of text documents. To search similar documents from the huge collection of document, a vocabulary is used which map each term given in the search query to the address of the corresponding inverted file; the inverted files are then read from the disk; and taking the intersection of the sets of documents for AND, OR, NOT operations, they are merged [1]. To help retrieval process, it requires additional structures such as document frequency, term frequency of each term in the given document.

A common problem of Information Retrieval is that users have to browse large number of documents containing both relevant and non relevant documents before finding similar documents. Clustering keeps relevant documents combined in a single group and hence fasten the process of Information Retrieval .There are many clustering algorithms available like K-means, Bisecting K-means, HFTC (Hierarchical Document clustering using Frequent Item sets), Hybrid PSO+K-means method and Global K-means [3],[5],[8].

Text summarization improves the speed required for processing the retrieval of documents [2]. Priority of documents is made based on the date or conclusion or the abstract provided by the authors of the document. However it is not possible as, not all documents come with an abstract.

5. PROPOSED METHODOLOGY
1. Document Preprocessing
Many document classification techniques are available in data mining. Data mining analyzes the data in the numerical form and this is the requirement of many algorithms. For applying these algorithms the text document has to be converted into numerical representation [10].

2. Feature selection
The main aim of feature selection is to choose a subset of input variables as per the requirements of user by eliminating features, which are not required or irrelevant. This increases the user friendliness, increases analytic accuracy, learning efficiency is also to be increased and reducing complexity of learned results. The main aim of feature selection is to find relevant feature subset that produces higher classification accuracy.

3. Making clusters
After feature selection we can make the clusters of the data which is having more correlation. So making clusters reduces the repletion of data and from one cluster you can choose the required value. The values are to be represented in the form of binary values 0 or 1. Using neural network you can make the clusters.
4. **Document classification**

In general, document classification means assigning documents to a fixed set of categories based on the requirements. Using neural network or weka tool one can classify the documents after training the dataset by selecting set of attributes.

![Fig 1. Flow Diagram of Document Mining](image)

6. **IMPLEMENTATION MODEL**

Our design prototype will consist of the following components:

1. A subsystem to automatically mine law firm data for each document. This data will be used to query and retrieve information. For example Law firms’ data would be computerized to our database. The documents will certainly be processed by classifying them on the basis of the date, type of document, priority based as per the requirement of the user. Thus we would have a real-world electronic database. One advantage of such a system is that it would be easy to find relevant laws and its related documents from vast amount of databases.

2. In order to create a specific, controllable database for research on querying and information retrieval of relevant information, it is proposed that a new electronic system be established. This system will allow the user to get the electronic version of the relevant information.

7. **CONCLUSION**

One of the major problems that the users face nowadays in getting the relevant information from the vast amount of database. In this paper we have presented a survey which covers techniques and methods that promise to enable us to mine documents information from vast amount of databases. Also we have proposed a system, an intelligent/information system, to satisfy the user requirements. Our framework would facilitate intelligent and fast retrieval of information as per the user’s requirement, such as getting relevant Law firm information fast. This paper focuses on the current issues in document processing.

**References**


AUTHOR
Ankita, Dhanashree, Diksha, Mangesh, Pranoti,
All are the students of final year (CSE), GNIET, Nagpur
Mrs. Reena Thakur has received a B.E.(CSE) degree from Amravati University and Master degree in Computer Science and Engineering from Uttar Pradesh Technical University Lucknow(U.P.). She is having 17 yrs of teaching experience. She has written three books. She has more than 10 publications to her credit in international journals, conferences as well as in IEEE Explore. Her fields of interest are Image Processing, Data Mining, and Computer Graphics. She is presently working in Guru Nanak Institute of Engineering and Technology, Nagpur.