TEXT MINING ON CRIMINAL DOCUMENTS

[1]RIYA, [2]NAMITA GANDOTRA

[1,2]Shoolini Univerity, Solan, H.P.
E-mail: [1]mrsmalhans@gmail.com, [2]namita.saini13@gmail.com

Abstract— Criminal cases that has similar facets means that case has many faces that may receive very different punishment in court sentence. Court system faces severe challenges on this problem, while search for similar criminal cases are manually from historical judgments database. It is very time consuming and it is not possible for a judge to do so for most of the criminal cases. This paper tries to use text mining technology to provide best solution for the criminal cases.

Index Terms— Criminal Case Judgments, (Referential Prison Term Generation, Judge's Sentence.

I. INTRODUCTION

The main idea of this paper is used for police investigation document of criminal case send from district police office , parse this document into keyword table, then match this data with the trained database in the form of key word table format of court's judgment. Using Cosine similarity algorithm of text mining technology to calculate coefficient of similarity, base on highest coefficient, we will find the closest judgment of the criminal case.

Text mining is also called as data text mining which is equal to text analytics which means deriving high quality information from raw data or raw text. It can be done by devising of patterns and trends through pattern learning. Text mining involves process of structuring the input data along some derived linguistic features and removal of others un-useful data or meaning-less data and then that data is inserted into the database and then deriving data with the structured data and the evaluation and the interpretation of the output.

Main Idea of This Research

The Procedure of criminal case handling in court. Procedure of drug abuse criminal case in court, most and large, start from receiving a Indictment send by District prosecutor's office. In indictment, facet of this criminal case, articles of law applied, list of evidences, information of offender as well as victims are described. Judge base on these materials, testimony of on the court, make his judgment. When making judgment, judge need to clarify the consistence of facet and evidences of criminal case, previous criminal record and if possible, check the term of punishment that similar case been sentenced, to decide which articles of criminal law be applied and to what should extent should the offender be punished. Security and safety are high priority concerned issues of civics in a country, court sentence of criminal cases, focused by press, bring strongly impact to citizens. People always request for fare punishment to those criminal case.

Security and safety always rank as high priority concerned issues in country. People always have a request for adequate and fare punishment to those criminal cases, while we may found most and large, criminal case that has similar faces may received very different punishment. This is the serious challenge to court system. To make fare and balanced judgment, judges must search for similar criminal cases from historical judgments database, to find out punishments that had been made before he makes his decision. In year 2008 each judge of district courts in Ludhiana and Chandigarh has 75% (average) per month. Now in year 2015 each judge of district court in LDH have 65% (average) per month, and in High court CHD each judge have 60% (average) per month complaint sounds on overloading is laud and clear. Searching for similar criminal cases manually from historical judgments database is very time consuming and is it not possible for a judge to do in most criminal cases handling except some critical cases. This paper will tries to use text mining approach to solve problems faced by judges and people should get more aware about the rules and punishments so that they can safe their children from many things.

Text Mining parse unstructured document into meaningful elements and used to execute further work as data mining technology do. In this paper we try to use cosine similarity approach of text mining to find similar judgments for judge to make fare and balanced judgment. And govt. also should get aware so that govt. can make strict rules so that people should fear from doing anything wrong.

Text mining approach may also be used to solve another tough issue that judges are suffered. The quantity of criminal cases in year 2015 in Ludhiana is up-to 405,357, while 60% of criminal cases are crimes of drug abuse, public danger, larceny and fraud, these types of criminal cases are relative simple that cases of killing, corruption etc., but judges still spend much of the costly time to handle these cases.

If we classify judgments of drug abuse criminal cases into several clusters base on the similar facet of criminal case and assign a judgment template for each cluster base when we building up training database for cosine similarity comparison. Then we can use cosine similarity comparison approach to find the nearest judgments for a target indictment that send from

district prosecutor's office. We can then find the best fit judgment template base on which cluster it belongs, by the way we can digest import stuff from indictment (feature extraction) and paste it to the template, a draft of court judgment may generated automatically. Judge only need to make necessary modification instead of writing a full judgment manually.

Constrained by time and resource, this paper will focus on the first issue, other topic is govt. should get more alert and make strict and more strict rules so that nobody can think of doing any wrong work that must be punishable.

II. OBJECTIVE OF THE STUDY

Based on all the above discussion some of the objective of my study is as following:

- Collection of data regarding criminal cases.
- Making text files from doc files.
- With the help of database making a decision supportive.

Criminal case that has similar facets may receive very different punishment in court sentence. Court system faces severe challenge on this problem, while search for similar criminal cases manually from historical judgments database, is very time consuming, and is impossible for a judge to do so for most of criminal cases. This paper tries to use text mining technology to provide solution.

We study the multi-criteria decision making problems in which the information about preference values is expressed in the form of uncertain linguistic information. We introduce some operational laws of uncertain linguistic variables and the concept of the ideal point of preference values with uncertain linguistic information, and develop an uncertain linguistic weighted averaging (ULWA) operator. An ideal -point-based approach to multi-criteria decision making with uncertain linguistic information is proposed. Finally, an illustrative numerical example is also given to verify the developed approach and to demonstrate its feasibility and practicality.

Among supplementary teaching methods, online teaching is the most important one. It is not limited to time and location, and can record the learning behaviors, such as online frequency, online time, and reading materials, in the learning portfolio. However, such recordings are mostly accumulative data, which cannot reflect the outcome of the, and provide appropriate learning aid. The Term Frequency Inverse Document Frequency (TFIDF) of keywords was treated as the criterion for the quality of the posted paper. Experimental results indicated that students who posted more non-useful articles have poorer learning outcome, and the quality of posted papers and learning outcome are closely correlated. Keywords: text mining, online teaching, subject categorization, supplementary learning.

Data mining is often used during the knowledge discovery process and is one of the most important subfields in knowledge management. Data mining aims to analyze a set of given data or information in order to identify novel and potentially useful patterns. These techniques such as decision trees, artificial neural networks, rule mining and genetic algorithms are used to discover patterns or knowledge that are previously unknown to the system and the users. Data mining has been used in many applications such as marketing, engineering, medicine, crime analysis, expert prediction, web mining, and mobile computing.

Text mining is variation on a field called data mining that tries to find interesting patterns from large databases. Text mining also known as intelligent text analysis, Text data mining refers to the process of extracting interesting and knowledge from unstructured text.

Text mining is young interdisciplinary field which draws on information retrieval, data mining, machine learning, statistics and computational linguistics. As most information is stored as text, text mining is believed to have a high commercial potential value.

Text mining is similar to data mining, except that data mining tools are designed to handle structured data from databases, but text mining can work with unstructured or semi-structured data sets such as emails, full-text documents and HTML files etc. As a result, text mining is a much better solution for companies. To date, however, most research and development efforts have centered on data mining efforts using structured data. The problem introduced by text mining is obvious: natural language was developed for humans to communicate with one another and to record information, and computers are a long way from comprehending natural language.

Humans have the ability to distinguish and apply linguistic patterns to text and humans can overcome obstacles that computers cannot easily handle such as slang, spelling variation and contextual meaning. However, although our language capabilities allow us to comprehend unstructured data, we lack the computer's ability to process text in large volumes or at high speeds.

Start, with a collection of documents, a text mining tool would retrieve a particular document and preprocess it by checking format and character sets. Then it would go through a text analysis phase, sometimes repeating techniques until information is extracted. The resulting information can be placed in management information system, yielding an abundant amount of knowledge for the user of that system.

The problem of Knowledge Discovery from Text is to extract explicit and implicit concepts and semantic relations between concepts using Natural Language Processing techniques. Its aim is to get insights into large quantities of text data. Knowledge Discovery from text draws on methods from statistics, machine learning, reasoning, information extraction,

knowledge management, and others for its discovery process. Knowledge Discovery from Text plays an increasingly significant role in emerging applications, such as Text Understanding.

Data mining tools can answer crime questions that have traditionally been too time consuming to resolve. They search databases for hidden patterns, finding critical information that experts may miss because it lies outside their expectations. The overall goal of the data mining process is to extract information from data set and transform it as an understandable structure for further use.

Text summarization is helpful for trying to figure out whether or not a lengthy document meets the user's need and is worth reading for further information. With large texts, text summarization software processes and summarizes the document in the time it would take the user to read the first paragraph. The key to summarization is to reduce the length and detail of a document while retaining its main points, overall meaning. The challenge is that, although computers are able to identify people, places and time, it is still difficult to teach software to analyze semantics and to interpret meaning.

III. METHODOLOGY

To achieve the objectives of research, the steps to be followed are as under:

Step1:-Firstly we have to collect the data which contains the text data in the form of word and pdf files. The major collection of the data is District Court Ludhiana and High court Chandigarh.

Step2:-We will read the data with the help of OCR

Step3:-After this we extract the major data fields from the digital data with the help of this tool.

Cosine similarity algorithm:-

There are many measures of similarity, shared Word count, word count and bonus and cosine similarity. The most obvious measure of similarity between documents is count of their words. For an information retrieval system, we likely to have a global dictionary, where all potential words will be included in a dictionary, with the exception of stop words. Every documents used as a training data or spreadsheet will be parsed into an entry of dictionary, showing as a vector.

The classical information retrieval approach to comparing documents is cosine similarity

Cosine similarity is a measure of similarity between two vectors of n dimensions by finding the cosine of the angle between them, often used to compare documents in text mining. Given two vectors of attributes with n dimensions, A and B, the cosine similarity, -, is represented using dot(.) product and magnitude as

Similarity=cos (□) =A.B/||A||B||.

For text matching, the attribute vectors A and B are usually the term frequency vectors of the documents. The cosine similarity can be seen as a method of normalizing document length during comparison.

The cosine similarity of two documents will range from 0 to 1, since the term frequencies cannot be negative. The angle between two term frequency vectors cannot be greater than 90°. 1 meaning exactly the same, with 0 indicating independence and in between values indicating intermediate similarity or dissimilarity.

IV. SELECTION OF TEXT MINING TOOL

English text (words) parsing tool have been well developed, tools are available at popular database management system software, while Chinese text parsing tool are limited and still at developing status. Chinese text is no blank to mark word boundaries as result, indentifying meaningful words is difficult, because of segmentation ambiguities and occurrence of unknown words, Chinese document parsing tool. Automatic Sentence Parsing and Semantic Composition under E-How Net (ASPSC) in short developed by Jiann. Researcher of the Institute of Information Science IIS was adapted in this research.

Since the facet and evidences of criminal case are listed in prosecutor's indictment, to save time for judges, can we identify? The major facet and list of evidences of indictment document and use those to compare with the judgments history of court, to find similar cases and articles of law as well as the term of punishment been applies in these cases? Use the indictment of criminal case that send from district prosecutor's office, parse this document into key words table (vector of words count that appear in document, table1 shows the example), then use this key word table1 (vector) to match with each vector of court's judgments built in trained database (i.e. dictionary, in key word table format, table2 shows the example).

Table 1: Determination of frequency

Key Words	Word1	Word2	Word 3	....	Word n
Counts of frequency	2	0	1		5

Table 2: Similarity between indictment and judgments

Key Words	Word 1	Word 2	Word 3	..	Word n	Cluster (Label)
Judgment1 word count	1	1	0		3	B
Judgment 2 word count	2	1	1		4	A
.....						
Judgment m	1	1	0		2	A

Using cosine Similarity algorithm of text mining technology to calculate coefficient of similarity between indictment and judgments. Base on the highest coefficient, we will find closest judgment of this criminal case.

Select highest 5 coefficients of judgments, a referential judgment report will be generated automatically, which provide information of similar criminal facets and term of punishments to judges for balance judgment. The quantity of judgments in referential report is adjustable. Owning to indictments of prosecutor is not open to public, hence; this research will use police's investigation documents instead (similar to indictment, police's investigation documents have information of criminal facet, evidences list, etc.) for training and testing key word dictionary.

Step4:- In the last step we will have the final data which can be used for judgment for the judge

The main idea of this paper is used for police investigation document of criminal case send from district police office , parse this document into keyword table, then match this data with the trained database in the form of key word table format of court's judgment. Using Cosine similarity algorithm of text mining technology to calculate coefficient of similarity, base on highest coefficient, we will find the closest judgment of the criminal case.

Security and safety always rank as high priority concerned issues in country. People always have a request for adequate and fare punishment to those criminal cases, while we may found most and large, criminal case that has similar faces may received very different punishment. This is the serious challenge to court system. To make fare and balanced judgment, judges must search for similar criminal cases from historical judgments database, to find out punishments that had been made before he makes his decision. In year 2008 each judge of district courts in Ludhiana and Chandigarh has 75% (average) per month. Now in year 2015 each judge of district court in LDH have 65% (average) per month, and in High court CHD each judge have 60% (average) per month complaint sounds on overloading is laud and clear. Searching for similar criminal cases manually from historical judgments database is very time consuming and is it not possible for a judge to do in most criminal cases handling except some critical cases. This paper will tries to use text mining approach to solve problems faced by judges and people should get more aware about the rules and punishments so that they can safe their children from many things.

If we classify judgments of drug abuse criminal cases into several clusters base on the similar facet of criminal case and assign a judgment template for each cluster base when we building up training database for cosine similarity comparison. Then we can use cosine similarity comparison approach to find the nearest judgments for a target indictment that send from district prosecutor's office. We can then find the best fit judgment template base on which cluster it belongs, by the way we can digest import stuff from indictment (feature extraction) and paste it to the template, a draft of court judgment may generated automatically. Judge only need to make necessary modification instead of writing a full judgment manually

- Constrained by time and resource, this paper will focus on the first issue, other topic is govt. should get more alert and make strict and more strict rules so that nobody can think of doing any wrong work that must be punishable.
- Collection of data regarding criminal cases.
- Making unstructured data from the structured data.
- With the help of that making a decision supportive and less time consuming.

V. SIMULATIONS AND EXPERIMENTAL RESULTS

The proposed solutions have been designed using Xilinx. The area-efficient carry select adder can also achieve an outstanding performance in power consumption. Power consumption can be greatly saved in our proposed area-efficient carry select adder because we only need one XOR gate and one INV gate in each summation operation as well as one AND gate and one OR gate in each carry-out operation after logic simplification and sharing partial circuit. Because of hardware sharing, we can also significantly reduce the occurring chance of glitch. Besides, the improvement of power consumption can be more obvious as the input bit number increases.

REFERENCES

[1] Agahi H, Mohammadpour A, Vaezpour M S. Predictive tools in data mining and k-means clustering: Universal Inequalities. 2013; 63:3:4 Issue 3-4;779-803.

[2] Bhushan J , Pushkar W, Shivaji K, Nikhil K .Searching Research Papers Using Clustering and Text Mining‖ International Journal of Emerging Technology and Advanced Engineering Website: www.ijetae.com, 2014; 4:4.

[3] Chauhan SR, Desai A. A Review on Knowledge Discovery using Text Classification Techniques in Text Mining". 2015; 111:6.

[4] Ghalib M, Vohra S, Vohra S, Juneja A. MINING ON CAR DATABASE EMPLOYING LEARNING AND CLUSTERING ALGORITHMS. International Journal of Engineering and Technology (IJET).

[5] Garrette D, Mielens J, Baldridge J. Real-World Semi-Supervised Learning of POS-Taggers for Low-Resource Languages. Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (ACL-2013). 583-592.

[6] Funmilayo K, Afolashade K, Oludele A. Text Mining Approach in Curtailing Cyber-Crimes in Nigeria. International Journal of Computer Science Trends and Technology (IJCST), 2015; 3:6.

[7] Ghosh S, Dubey S K. Comparative Analysis of K-Means and Fuzzy C-Means Algorithms. International Journal of Advanced Computer Science and Applications, 2013; 4:4.

★★★