PACIS 2010 Proceedings

Pacific Asia Conference on Information Systems (PACIS)

2010

Use Text Mining Approach to Generate the Draft of Indictment for Prosecutor

Chuan-hsi Chen
National Chengchi University, cchen7@mail.moj.gov.tw

Jeffery Y. P. Chi
National Chengchi University, ypchi@mis.nccy.edu.tw

USE TEXT MINING APPROACH TO GENERATE THE DRAFT OF INDICTMENT FOR PROSECUTOR

Chuan-hsi Chen, Department of Management Information Systems, National Chengchi University, Taipei, Taiwan R.O.C.,cchen7@mail.moj.gov.tw

Jeffery Y. P. Chi, Department of Management Information Systems, National Chengchi University, Taipei, Taiwan R.O.C.,ypchi@mis.nccy.edu.tw

Abstract

Motivation: The quantity of criminal cases year 2009 in Taiwan is up to 1.8 millions, Each prosecutor must handle over 211 cases per month, complaints on over loading is laud and clear. While 70 % of criminal cases are drug Abuse, public danger, larceny and fraud, these types of criminal cases may have different story though, the complexity are relative simple than cases of killing, corruption etc., but prosecutors still spend costly time on these cases handling. In this paper we try to use text mining technology to provide solution on this issue.

Approach: We use the police's investigation document of criminal case to compare with judgment history of court, and use Cosine Similarity algorithm to calculate coefficient of similarity, base on the highest coefficient, we find the closest judgment of this type of criminal case, that can be used to decide and generate the draft of indictment for prosecutor.

Key words: Text Mining, Clustering, Cosine Similarity, Criminal Case Judgment, Drug Abuse, indictment, prosecution

1. INTRODUCTION

Prosecutor is very expensive and limited resource of a country. To make best use of prosecutor’s resource, new IT technologies should be taken into consideration. The quantity of criminal cases of year 2009 in Taiwan is up to 1,899,851 (Ministry of Justice, 2010). Each prosecutor must handle over 211 cases per month. Complaint sounds on over loading; appear in most records of performance review meeting in prosecutor’s office (Prosecutor Office of High court, 2010). This situation may hurt the quality of case prosecution; even more, hurt the trust from citizen.

Minister of Justice regard this issue as high priority task to solve; while the total head count of prosecutor in Taiwan is limited by law, alternative approaches must be taken. Statistic data of Ministry of Justice shows that 70 % of criminal cases are crimes of Drug Abuse, Public Danger, Larceny and Fraud (Ministry of Justice, 2010), these types of criminal cases may have different story though, the complexity are relative simple than cases of money laundry, killing, corruption etc., but prosecutor still spend lot of costly time on these cases handling.

Text mining is an interdisciplinary field which combine several IT technologies of information retrieval, data mining, machine learning, statistics, and computational linguistics. Most information, common estimates over 80% (Seth Grimes, 2010), is currently stored as text, text mining is believed to have a high commercial potential value. Increasing interest is being paid to multilingual data mining: the ability to gain information across languages and cluster similar items from different linguistic sources according to their meaning (Wikipedia1, Feb. 2010). 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 technology to find solution for this issue.

Drug abuse is the majority of criminal cases in prosecutor’s office, 40 % of prisons in correction agencies (jail, detention house) are drug abuse offenders, if we can save time for prosecutor in handling this type of criminal cases, the effectiveness will be significant. This paper try to identify the major facet from investigation document of drug abuse case that send from district police office, and use Cosine Similarity method to find the closest judgment with similar facet, from historical judgments database. Then, we generate the draft of prosecution document (indictment) for prosecutor automatically. Prosecutor can just review indictment draft and make the necessary revise, instead of making full indictment manually. This approach may save a lot of costly time for prosecutors, and allow them focus on complex cases handling.

2. LITERATURE REVIEW

2.1 The Meaning of Text Mining.

Text mining is a process that analysis, edit, organize a set of documents, to find hidden feature, extract information, for further processing (Sullivan, 2001). The frequency of a specific phase that appear in a document, reveal the degree of importance of that phase. While, a specific phase appear in many documents (document frequency) also shows another important message (Salton, Buckley, 1988). Weiss comment that “Do we have a shortage of data? Not very likely, big data are available for further analysis”. Text represents over 80% of all information

handled by an organization (Seth Grimes,2010). Data mining is a mature technology, this method process structured numerical information. Text are often described as unstructured information, but if parse document into spread sheet formation, text and document can be transforms into measured value, and methods that used on programming and data mining can be applied in text mining area (Weiss, et. al., 2005).

2.2 The major functions of text mining.

The major methods that can be used in data mining can also be used in text mining (Sholom

M. Weiss, et al., 2005):

1. Document Classification: Once the text are transformed to the usual numerical spreadsheet format, documents can be organized into folders, one folder for each topics
2. Information Retrieval: A basic concept for information retrieval is measuring similarity: a comparison is made between two documents, measuring how similar the documents are.
3. Clustering: To deal with a set of documents with no known structure. The clustering process is equivalent to assign the labels needed for text categorization by way of numeric method instead of predefined categories by human.
4. Information Extraction: Parse text as well as numeral data from document into spreadsheet format, and extract the needed fields for further use.
5. Prediction: Ultimate goal of text mining is prediction, projecting from a sample of prior examples to new unseen examples.

2.3 Application of Text Mining

Text mining equipped with the ability to gain information across languages and cluster similar items from different linguistic sources according to their meaning, can be applied in wild range of applications (Wikipedia1, Feb. 2010):

- Security: Many text mining software packages are marketed towards security applications, particularly analysis of text sources with security concerns.
- Biomedical: A range of text mining applications in the biomedical literature has been described (K. Bretonnel Cohen & Lawrence Hunter, 2008). In the United States, the School of Information at University of California, Berkeley is developing a program called BioText to assist bioscience researchers in text mining and analysis
- Software and applications: Research and development departments of major companies, including IBM and Microsoft, are researching text mining techniques and developing programs to further automate the mining and analysis processes.
- Online Media: Text mining is being used by large media companies, such as the Tribune Company, to disambiguate information and to provide readers with greater search experiences, which in turn increases site "stickiness" and revenue.
- Marketing: Text mining is starting to be used in marketing as well, more specifically in analytical Customer relationship management.
- Sentiment analysis: for example, involve analysis of movie reviews for estimating how favorable a review is for a movie (Bo Pang, et. al, 2002).
- Academic: Text mining is one of import approaches to publishers who hold large databases of information requiring indexing for retrieval. This is particularly true in scientific disciplines, in which highly specific information is often contained within written text.

Text mining applied in automatic generation of prosecutor's indictment still not found in research literature.

2.4 Cosine Similarity

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 a 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 (Weiss, et. al., 2005).

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, θ, is represented using a dot product and magnitude as:

$$\text{similarity} = \cos(\theta) = \frac{A \cdot B}{\|A\|\|B\|}.$$

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 (Wikipedia2, 2010).

2.5 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 a result, identifying meaningful words is difficult, because of segmentation ambiguities and occurrences of unknown words (Chen, Keh-Jiann, 2002). Chinese document parsing tool: "Automatic Sentence Parsing and Semantic Composition under E-How Net (ASPSC in short)" Developed by Chen, Keh-Jiann. Researcher of The Institute of Information Science (IIS), was adapted in this research.

3. MAIN IDEA OF RESEARCH

3.1 The Procedure of Criminal Case Prosecution

The procedure of drug case prosecution in Prosecutor's Office, most and large, start from receiving a "Investigation Document" submitted by district police office. In investigation document, facet of this criminal case, evidences list and basic information of defendants as well as victims are described. Prosecutor base on these material, take further investigating actions if necessary or make indictment directly. When making indictment, prosecutor need to clarify the facet of this criminal case, check the basic information of defendants/victims include previous criminal record and to decide which articles of criminal law be applied and to what extent should the victims be accused.

3.2 Main Idea

Since the facet and evidences of criminal case are listed in police's investigation document, to save time for prosecutors, can we identify the major facet of investigation document and use this part of text to compare with the judgments history of court, to find the similar cases and the articles of law as well as the term of punishment been applied in these cases? In this research, we use Cosine Similarity algorithm to compare the facet /evidences of police's investigation document with judgments history of court by calculate coefficient of similarity. Base on the highest coefficient; we will find the closest judgment of this criminal case.

We parse the investigation document into key words table (vector of words count that appear in document, table1 shows the example), every new key word will be list in table, and set the counter of frequency as 1, repeated key word appear, add 1 to that word count.

Before we can use the cosine similarity method to make comparison, a trained database of judgment history must be build. Each judgment selected as a training case in this trained database, will be assigned a specified cluster (label), this cluster have similar facet and applied similar articles of criminal law and term of punishment. Every judgment were parsed into key word table, all key word tables consist into a trained database (i.e. dictionary, in key word table format, table 2 shows the example). The detail procedure to build trained database will be discuss in next section. Since we can find the nearest judgments for a target investigation document, we can then find a best fit indictment template for this criminal case base on which cluster it belongs. By the way, we can digest the needed stuff from police's investigation document (e.g. the facet of criminal case, evidence list, etc.) and insert into the chosen temple of indictment, then, a draft of indictment will be generated automatically. Prosecutor can just review this draft and make the necessary revise, instead of making full indictment manually. This approach may save much time for prosecutor and allow them focus on complex cases handling.

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

Table 1. Example of Key Word Table of Investigation Document

Key Words	Word1	Word2	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	B

Table 2. Example of Trained Database of Judgment (Dictionary)

4. RESEARCH METHOD/PROCEDURE

Lab experiment is used in this paper; "Drug abuse" is selected as the target area of criminal cases in this research. Details of research method described as follows:

4.1 Data / Tool Preparation

We choice 50 cases of drug-abuse judgments sequentially from the judgment database of Taipei District Court of year 2007(Justice Yuan, 2009) and parse it one by one, to build training database (dictionary) for further similarity comparison. Cases of these judgments are classified into two drug abuse commitment cluster (low and high) base on degree of punishment of court sentence. 39 cases are assigned to cluster of low; 11cases are assigned to cluster of high. 8 police's investigation documents are used for training and testing.

4.2 Build training database (Dictionary)

4.2.1 Creation

Use 50 judgments as entries of training database. Building procedures are:

1. Read first judgment, use parsing tool ASPSC to parse into keywords
2. Put first key word into first column of keyword table (word 1) and set the initial word count frequency as 1.
3. Pick second key word of judgment; compare it with first keyword (word1). If same, add 1 to word count of word1. If different, add new column of keyword (word 2) to keyword table, and set the initial word count frequency as 1.
4. Continue to pick next key word of first judgment until all key word are processed.
5. Read next judgment, repeat action 1-4, until all judgment had been processed.

4.2.2 First Phase Modification

When training database had been built, we found over hundred of key words been identified. Some of key words are cut inadequately by parse tool into two or tree single words, for example " the second level drug amphetamines" was cut into "the second", "level", "drug", "amphetamines", the meaningful parsing should be "the second level drug", "amphetamines". Hence, field expert (prosecutor) was involved to review keyword table, recombine words into meaningful key words, and delete some of keywords that don't any contribution to drug judgment.

The quantity of drug been possessed or used by suspects is a very important information for prosecutor or judge to decide the degree of punishment, and article of law be applied, but quantity are different in most of drug abuse cases, if we select different quantity as a key word, say 5 gram, 7gram, 12gram….., and put these key word into keyword table, the frequency of word count will be low, hence, will dilute the inference of this key word, and mislead that result.

4.3 Calculate Cosine Similarity

Parse new criminal case investigation document that submitted by district police's office into key word table. Only key words and its frequencies specified in dictionary will be used to build vector. Use this criminal case key word vector to compare with the judgment training database. Calculate coefficient of similarity by cosine similarity algorism, base on the highest coefficient, we will find the closest judgment of this criminal case. Sample output of cosine similarity illustrated at Table 3.

Investigation Document	Similar Court Judgments & drug commit level (cluster)	Term of punishment	Coefficient of Cosine Similarity
Case NO: 9632103800 Send by: Taipei Police Office Term of punishment: 7 months (real sentence from court) Drug commitment cluster: high	J01, cluster high	8 months	0.808452
	J02, cluster high	7 months	0.793051
	J03, cluster high	8 months	0.780720
	J04, cluster low	6 months	0.738548
	J05, cluster high	7 months	0.727392

Table 3. Sample Output of Cosine Similarity

4.4 Test and Modify (second phase) training database

8 police's investigation document come with its court sentence (term of punishment) were used to test and modify keywords in training database. Irritated testing and modification was taken place to find the highest similarity with right drug commitment cluster. Determination rule of a right match is: If investigation document belong to low cluster, then the coefficient of judgment with low drug commit cluster must higher than judgments with high drug commit cluster, and vice versa. The illustration of determination rule can also be found in table 3.

5. FINDING OF THIS PAPER

After repeating testing and modification, we experience an acceptable result in this experiment. Only one out of eight police's investigation documents match with wrong judgment that has different drug commit cluster, 87.5% of accuracy is reached. Table 4 shows the result of this experiment. If we classify judgments of drug abuse cases into detail clusters match to real work need, the idea of this paper can be practice. This can be topics for further study.

Second finding in this research is: too many key words in dictionary may not increase the effectiveness of cosine similarity comparison, on contrary, it will dilute the influence of the important keywords and thus make cosine coefficient appear poor performance.

Third finding is: the quantity of drug in investigation document is important for prosecutor/judge to decide the term of punishment in drug abuse case, this type of parsed data is not effective for cosine similarity comparison, but it can be use as a structure data for programming processing, and making more precise decision.

CASE	Investigation Documents	Best Coefficient of Similar Judgment	Second best Coefficient of Similar Judgment	Drug Commit Cluster matched
1	Case NO: 9630970100	Judgment NO: J96939 0.960276	NO: J97252 0.917746	YES
2	Case NO: 950008194	NO: J961563 0.945905	NO: J9744 0,878833	YES
3	Case NO: 9632310700	NO: J972234 0.992277	NO: J952658 0.979798	YES
4	Case NO: 9630077000	NO: J97570 0.842700	NO: J971128 0.834057	YES
5	Case NO: 9632103800	NO: J972 0.808452	NO: J9774 0.793051	NO
6	Case NO: 9632178400	NO: J97278 0.938350	NO: J972632 0.918262	YES
7	Case NO: 930022576	NO: J97278 0.818756	NO: J97134 0.808122	YES
8	9 Case NO: 300001153	NO: J97120 0.898026	NO: J9734 0.894427	YES

Table 4. Cosine Similarity Comparison of Investigation Document and Judgment

6. CONCLUSION

Base on the experiment result in this research, cosine similarity comparison approach has been found effective for finding the similar judgment to make the draft of prosecution indictment when received a drug abuse criminal case send from police office. This approach can save costly time for prosecutors, and allow them focus on complex cases handling. Larger scale experiment is suggested to execute to prove complex clustering (match with the real work's need) is performing well also. Owing to the decision type of drug abuse case is not so complicated, it is possible for researchers using decision tree approach to find best match of court judgment, base on the facet of criminal case. This can be another topic for further research.

References

Berry, M. J. A. and Linoff, G.. (1997). "Data Mining Technique for Marketing, Sale, and Customer Support", New York: John Wiley Computer.

Bo Pang, Lillian Lee, and Shivakumar Vaithyanathan (2002). "Thumbs up? Sentiment Classification using Machine Learning Techniques". Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP). pp. 79–86.

Deogun, J.S.; Sever, H.; Raghavan, V.V.(1998). "Structural Abstractions of Hypertext Documents for Web-based Retrieval", Proceeding of Ninth International Workshop on Database and Expert Systems Applications,pp.385 -390.

IBM, (1998). "Intelligent Miner for Text: Getting Started", IBM Cor.

K. Bretonnel Cohen & Lawrence Hunter (January 2008). "Getting Started in Text Mining". PLoS Computational Biology 4 (1): e20.

K. J. Chen, W. Y. Ma (2002). "Unknown Word Extraction for Chinese Documents," Proceedings of COLING, ,pp. 169~175,

Ministry of Justice,(Jan.2010), "Statistic Digest of Justice Affair", p.2, p.12

Prosecutor's Office of High Court, Ministry of justice, Mar. 2010, "Memorandum of Prosecutor General's Meeting 2010", p3.

Salton, G.& C. Buckley,(1988). "Term Weighting Approaches in Automatic Information Retrieval,"Journal of Information Proceeding and Management, Vol.24, No. 3, p.513-524.

Sholom M. Weiss, et al., (2005). "TEXT MINING—Predictive Methods for Analyzing Unstructured Information" Springer P.7

Sullivan, D. 2001 "Document Warehousing and Text Mining", Wiley Computer Publishing, pp. 326

Reference from Internet :

Ananyan S. "Text Mining Application and technology", ,http://www.vbits98.com.

Justice Yuan,(2008).. http://www.judicial.gov.tw/juds/index1.htm

The Institute of Information Science (IIS), Academia Sinica, http://ckipsvr.iis.sinica.edu.tw

Wikipedia1, Feb. 2010, http://en.wikipedia.org/wiki/Text_mining

Wikipedia2, Feb. 2010,http://en.wikipedia.org/wiki/Text_mining#Security_applications

Seth Grimes, 2010, http://www.clarabridge.com/default.aspx?tabid=137&ModuleID=635&ArticleID=551