ISSN: 2321-7782 (Online)
Volume 2, Issue 12, December 2014
International Journal of Advance Research in Computer Science and Management Studies
Research Article / Survey Paper / Case Study
Available online at: www.ijarcsms.com

Feature Extraction Based Legal Document Summarization

S. Santhana Megala[1]
Research Scholar
PRIST University
Thanjavur, Tamil Nadu – India

Dr. A. Kavitha[1]
Dept. of Computer Science
Kongunadu Arts & Science College
Coimbatore, Tamil Nadu – India

Dr. A. Marimuthu[3]
Dept. of Computer Science
Govt. Arts College
Coimbatore, Tamil Nadu – India

Abstract: ***Basically Legal Documents were little bit tough to understand and it is too long, hence a headnote, a brief summary of the Legal Document, is a needed one in the current scenario. In recent days there are number of research work is going on to automate this Head note preparation process. The generation of headnote is a time consuming process for Advocates as well as Judges for Arguments and for Decision Making. The main drawback in the legal field is that, they were not coherent and can't able to convey the relative relevance of the case. In this paper, Head note is prepared automatically from the legal document. A set of legal documents were extracted from the Website, and used as training data. The present work implements Fuzzy Logic Technique to generate the Headnote summary.***

Keywords: Fuzzy Logic, Legal Document, Sentence Scoring, Feature Extraction, Text Summarization.

I. INTRODUCTION

Information Overloading is a main problem in the digitized world due to the increasing amount of digital data, which leads to the invention of new technological tool to create head note summary. Creating a Headnote is a process of comprising the large amount of text data into a minimum amount in the digital documents. Text Summarization plays a vibrant role in creating the summary by extracting the important text information's in the document. Text Summarization helps the people to rapidly understand huge amount of information within a limited time, which invites the technological people to do more research in this field, to improve this novel systems to the next level.

Ultimately the text summarization process is categorized into two types, one is abstraction and other is extraction. Abstraction constructs the summary, which acts as a supplementary for the original document i.e it employs the words and phrases which is not present in the original document, it changes the word with same meaning. On the other hand, Extraction process creates the summary by selecting important sentences from the original document. But maximum research work is done by using the extraction techniques because there are more number of research work was done using extraction techniques, which is a simple and easy one.

In Automatic summarization, the sentence extraction features were used to make the job easy for Fuzzy Logic to get more accuracy in the Summary Generation. In this approach 11 feature Extraction techniques were used such as Sentence Position, Proper Noun, Sentence Length, Term Frequency – Inverse Sentence Frequency (tf * isf), Sentence to sentence similarity, Citation, Local Features and Layout Features, Paragraph Structure, Thematic Word, Indicators / Cue Phrases, Legal Thesaurus.

This paper established in cue with a related works done by the researchers in the legal field in section (II), and Feature Extraction Techniques were discussed in the section (III). In Section (IV) the most popular Extraction Based Text Summarization Method using Fuzzy Logic were discussed. The Experiments and Results were covered in section (V). Finally the conclusion is noted in the section (VI) and plan for future work is discussed in section (VII).

II. Related Work

The input Legal Document is first segmented into thematic structure based on the words present in the document, after that only filtering, selection, and Production was done within the segmented region in [12]. A Topic Based Summarization, was done in [9] using LDA algorithm, which identifies the important topics in the Legal Documents. An Automatic text Summarization based on feature extraction techniques using Fuzzy Logic on news documents were used in [4].

III. Feature Extraction Techniques

A. *Pre-Processing*

The pre- processing step consists of cleaning the noisy text covering grammatical and typographical errors. The foremost problem in text summarization is that the size of the document is not well-known. Thus each word in the documents was denoted by the terms in the vector space model, which grounds the number of dimensions as too high for the text summarization algorithms. This pre- processing method plays a dynamic role in reducing the number of dimensions handed by the text summarization process. In this paper, below mentioned pre-processing methods were applied, namely, Case Folding, Stop Word Removal, Punctuation Removal, White Space Removal, Word Stemming, Key Phrase Identification, Sentence Segmentation and Tokenization.

A process of translating all the characters to the small characters is called Case Folding, like "ACT", "Act" to "act". The process of removing the insignificant words which appears repeatedly in the document and provides not as much of meaning in the text processing is called Stop Word Removal. The process of removing the all undesirable punctuations from the document except dot (.) operator, which act as a sentence separator, is called Punctuation Removal.

In Legal Document, the extra white space is inserted for some formatting purpose, which occupy more spaces, Removal of Extra White spaces is the process of eliminating the additional white spaces, which cuts the size of the document. A method is proposed for Word Stemming, which is the process of producing the root word, by removing the suffixes and prefixes of each word in the document. Identification of important phrases using relative frequency approach, by finding the occurrence of the word pairs is called Key Phrase Identification. The process of identifying and separating the paragraph into sentence is called Sentence Segmentation. The process of splitting the Sentence into individual words is called Tokenization.

B. *Sentence Feature Extraction*

Title After completing the Pre- processing step, by using feature extraction technique, each sentence in the document is denoted by a vector point. In this paper 11 feature extraction techniques were used, which gives a value stuck between "0" to "1". The 11 features were as follows:

- **Sentence Position**

In a Legal Document, the Sentences occurring in the first paragraph as well as sentences occurring in the last paragraph may contain important information's. The following formula is used in order to give high score to those sentences.

$$f_1 = \begin{cases} 1 & for \quad 1^{st} \text{ and } n^{th} \text{ Line} \\ 0.8 & for \quad 2^{nd} \text{ and } n-1^{th} \text{ Line} \\ 0.6 & for \quad 3^{rd} \text{ and } n-2^{th} \text{ Line} \\ 0.4 & for \quad 4^{th} \text{ and } n-3^{th} \text{ Line} \\ 0.2 & for \quad 5^{th} \text{ and } n-4^{th} \text{ Line} \\ 0 & Otherwise \end{cases} \quad \ldots\ldots\ldots equ\ (1)$$

❖ **Proper Noun**

The sentence, which contains more Named Entity called as Proper Noun, which is the important sentence that should be included in the Document Summary. A Named Entity should be started with a Capital Letter only, based on this we can find the Proper Noun in a sentence.

$$f_2 = \frac{No.of\ Proper\ Noun\ in\ a\ Sentence(S_i)}{Sentence\ Length\ (S_i)} \quad \ldots\ldots\ equ(2)$$

❖ **Sentence Length**

Sentence Length is a measure which is used in identifying the best sentences for summary. The number of words in a sentence is counted, first and then it is normailized to get the Length of the sentence.

$$f_3 = \frac{No.\ of\ Words\ Occuring\ in\ a\ Sentence(S_i)}{No.\ of\ Words\ Occuring\ in\ the\ Longest\ Sentence} \quad \ldots\ldots\ equ(3)$$

❖ **Tf * isf (Term Frequency – Inverse Sentence Frequency)**

Term Weight is used to calculate the importance of sentence by finding the frequency of occurrences of the term within a document, which is also called as raw term frequency. But it is not the case, because importance does not increase proportionately with raw term frequency. So, there is a need in the inverse sentence frequency, which filter the important word that occurring in the sentences.

$$f_4 = \sum_{1}^{n} tf * isf \quad \ldots\ldots\ equ(4)$$

Where n = No. of Words in the Sentence

$$tf = \frac{No.of\ times\ term\ t\ appears\ in\ a\ document}{Total\ No.of\ terms\ in\ the\ document} \qquad isf = \log\left(\frac{Total\ No.of\ Sentences\ in\ the\ document}{No.of\ Sentence\ in\ which\ term\ t\ occurs}\right)$$

❖ **Sentence to Sentence Similarity**

Similarity to Neighbouring Sentence finds the similarity between each sentence in the document, which is computed by using the cosine similarity measure, which compares all the sentences in the document.

$$f_5 = \cos(\theta) = \frac{A.B}{|A||B|} = \frac{\sum_{i=1}^{n} A_i\ x\ B_i}{\sqrt{\sum_{i=1}^{n}(A_i)^2}\ x\sqrt{\sum_{i=1}^{n}(B_i)^2}} \quad \ldots\ldots\ equ(5)$$

❖ **Citation**

Citation Denotes referring someone, which is a needed one in the legal field. Because Indian Legal System follows the Common Law System where, the judgements will referred by some other cases for Arguments and for Judgement. The citation is identified by the keyword v. or vs.

$$f_6 = \begin{cases} 1 & if\ in\ **[**\ v./vs.\ **]\ or\ **\ v.**\ or\ **vs.** \\ 0 & Otherwise \end{cases} \quad \ldots\ldots\ equ(6)$$

❖ **Local Features & Layout Features**

Each Legal Document will have a layout structure, from that the data related to the head note can be obtained, like, Judge Name, Court Name, Date, Petitioner & Respondent name etc.. The Local Features & Layout Features will get all the details related to the head note of the Legal Document.

$$f_7 = \begin{cases} 1 & \text{if } \textit{In the High Court} ** / \textit{Dated} / \textit{The Hon'ble} / \textit{The Honourable} \\ & \textit{W.P/Writ Petition /Petitioner/ Respondent} \\ 0 & \textit{Otherwise} \end{cases} \quad \ldots\ldots\ldots equ\,(7)$$

❖ **Paragraph Structure**

Every Document has a unique internal structure of the Paragraph, which have the high level sum-up in the starting as well as in the ending. Paragraph Structure will assign high score to the first and last paragraph.

$$f_8 = \begin{cases} 1 & \text{if } 2.\ (\textit{Occurs after the heading ORDER} / \textit{Common Order} \\ & **.\ (\textit{Para before the word TO appears} \\ 0 & \textit{Otherwise} \end{cases} \quad \ldots\ldots\ldots equ\,(8)$$

❖ **Thematic Word**

In Legal Document the final decision is the important one. The word which denotes the main theme of the document is mentioned by the Thematic Word. If the word present in a sentence then that sentence is an important one, it should come in the summary.

$$f_9 = \begin{cases} 1 & \text{if } \textit{dismiss, dismissed, allowed, Partly allowed, disposed of,} \\ & \textit{order accordingly, ruled discharged, remitted back,} \\ & \textit{reference answered in positive / negative, no costs} \\ & \textit{no merits} \\ 0 & \textit{Otherwise} \end{cases} \quad \ldots\ldots\ldots equ\,(9)$$

❖ **Indicators/Cue Phrases**

Cue Phrase denotes frequently used key phrases, which acts as the indicators of expressing the fact finding of the case as well as the judge. The indicators act as an important one in finding the rhetorical roles in the summary.

$$f_{10} = \begin{cases} 1 & \text{if } \textit{Second appearing, Question for consideration, no reason,} \\ & \textit{points to be considered, pertinent to note, we do not find,} \\ & \textit{relied on the decision, counsel submitted, appeal, suit, in my view} \\ & \textit{we therefore answer the question, revision, review, we find} \\ & \textit{case before us, impunged, quashed, Bone of Contention} \\ & \textit{in the light of , learned counsel, Contention contends, issue} \\ & \textit{points for consideration, we found, we agree,} \\ 0 & \textit{Otherwise} \end{cases} \quad \ldots\ldots\ldots equ\,(10)$$

❖ **Legal Thesaurus**

Legal Thesaurus means words or phrases that include the Legal words of basic vocabularies from a training data.

$$f_{11} = \begin{cases} 1 & \text{if } \textit{Appointment, Recuritment, Suspension, Termination, Misconduct} \\ & \textit{Disciplanary Proceeding, Probation, Service Conditions, Super Anuation,} \\ & \textit{Transfer, Reinstatement, Reversion, Allowances, Detuction, Graduity} \\ 0 & \textit{Otherwise} \end{cases} \quad \ldots\ldots\ldots equ\,(10)$$

IV. EXTRACTION BASED TEXT SUMMARIZATION METHOD

a. *Text Summarization Based on Fuzzy Logic*

To Implement the Legal Text Summarization using Fuzzy Logic, First the Feature Extraction process has to complete for all the 11 features. Fuzzy Logic system design involves in selecting membership function and fuzzy rules. The action of the fuzzy logic system will openly affect by the selection of fuzzy rules and membership functions. The four main components of the Fuzzy Logic System were: Fuzzifier, Inference Engine, Defuzzifier, and the Fuzzy Knowledge Base. In the fuzzifier section, the membership function translates the inputs into linguistic values.

The inference engine refers the Fuzzy rule base, which contains fuzzy IF THEN rule to derive the linguistic values. At last, the Defuzzifier converts the linguistic variables to the final crisp values from the inference Engine, using output membership function. The final sentence score was derived. In the Defuzzfication step, the output membership function step is divided into three membership functions, namely: Unimportant, Average, Important, Which converts the result of the inference engine into a crisp output to obtain a final sentence score for each sentence.

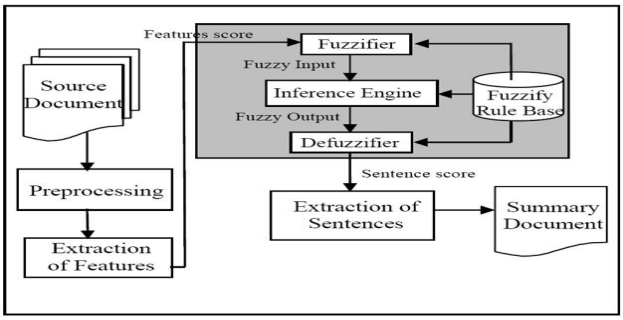


Figure 1: Shows text summarization based on fuzzy logic system architecture.

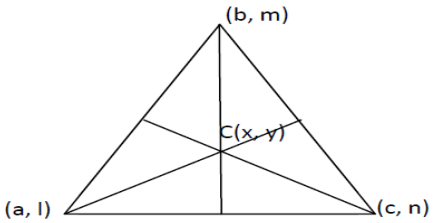


Figure 2: Fuzzy Centroid Calculation

The Membership functions used in the Fuzzy logic was based on the fuzzy centroid method, which calculates the score for the entire sentences present in the legal document. Fuzzy Centroid method used generalized triangular membership function, to obtain the sentence score, which depends on the three parameters 'a', 'b' and 'c'. In which the position of the parameters 'a' and 'c' are left and right most feet of a triangle and 'b' is the peak of a triangle.

The output value was obtained from zero to one for each sentence, based upon the sentence features and knowledge base. The above said value shows the degree of importance of the sentences that present in the final summary. The formula to calculate the fuzzy centroid (1) is given below.

$$C(x,y) = \left(\frac{a+b+c}{3}, \frac{l+m+n}{3}\right) \quad \text{--------- (equ 12)}$$

The values a, b, c were the standard values of Low Medium and High, respectively and the values l, m, n were the calculated values of Low Medium and High, respectively. Defining IF-Then rules is the important one in the Inference Engine. Sample IF –Then rules for our feature extraction measures was mentioned below.

*IF (Sentence Position is VH) and (Proper Noun is H) and (Sentence Length is VH) and (tf*isf is H) and (Sentence to sentence Similarity is VH) and (Citation is H) and (Local & Layout Features is VH) and (Paragraph Structures is H) and (Thematic Word is H) and (Indicators/Cue Phrases is M) and (Legal Thesaurus is H) THEN (Sentence is important)*

V. EXPERIMENTS AND RESULTS

In this paper, the system is developed using Fuzzy Logic Method and implemented using c#. The experimentation is done by taking 150 Legal Documents from the legal website, among that 50 Documents from the Service Law, 50 Documents from

Industry Law, and 50 Documents from Constitutional Law. On the given data set, First pre-processing technique is applied, followed by the feature extraction techniques, which obtains the feature Score for the sentences present in the Legal Document.

TABLE 1

SAMPLE FEATURE SCORE FOR THE LEGAL DOCUMENT

D	Feature Score										
	f1	f2	f3	f4	f5	f6	f7	f8	f9	f10	f11
S1	0.4	0.2	1	0.9	0.2	0	1	0	1	0	1
S2	0.2	0	0	0.5	0	0	1	0	0	0	0
S3	0	0	0	0	0.8	1	0	1	0	0	0
S4	0	0	0	1	0	1	0	0	0	1	0
S5	0	1	0.7	0	0	0	0	0	1	0	1

TABLE 2

SAMPLE OUTPUT FOR JUDGMENT SUMARIZATION

IN THE HIGH COURT OF DELHI AT NEW DELHI. 14th December, 2011. ORIENTAL INSURANCE CO. LTD - Appellant. KAUSHALYA DEVI & ORS- Respondents. HON'BLE MR. JUSTICE G.P.MITTAL. The Appellant Oriental Insurance Company Limited is aggrieved by an award dated 27.08.2004 whereby it was made liable to pay the compensation in the seven Claim Petitions decided by a common order with a right of recovery from Vinod Kumar . The Tribunal by the impugned award found that the accident took place on account of rashness and negligence on the part of the tempo driver Ram Pal. It was submitted that the persons were travelling in the tempo in violation of the terms of insurance policy and thus, it is not liable to pay the compensation. New India Insurance Company vs. Asha Rani 2003 ACJ (1) SC. National Insurance Company Ltd. vs. Baljeet Kaur & Others 2004 ACJ 428. It is hereby held that respondent no.3 is liable to pay the compensation to the petitioners. The question which thus, arises is whether the respondent no.3 is entitled to recover the said compensation from the owner of the tempo i.e. respondent no.2. It has been held in the case of Sawaran Singh Vs. National Insurance Company 2004. In Skandla Insurance Company Ltd. Vs. Kokilaben Chandravadan & Ors. 1987. It is not entitled to recover the compensation amount from respondent no.2.". Section 147 of the 1988 Act did not impose any statutory liability on the owner of vehicle to get his vehicle insured for any passengers travelling in a goods vehicle and therefore, the insurers would not be liable. Sections 2(8), 2(25), 2(29) and 2(33) of the 1939 Act . Sections 2(14), 2(35), 2(40) and 2(47) of 1988 Act. Mallawwa (Smt.) and Ors. v. Oriental Insurance Company Ltd. and Ors. (1999) 1 SCC 403. Section 2(35) of 1988 Act does not include passengers in goods carriage whereas Section 2(25). we are of the opinion that as the provisions thereof do not enjoin any statutory liability on the owner of a vehicle to get his vehicle insured for any passenger travelling in a goods vehicle, the insurers would not be liable therefor." in National Insurance Company v. Baljit Kaur & Ors. (2004) 2 SCC 1. New India Assurance Company v. Satpal Singh (2000) 1 SCC 237. in New India Assurance Company Limited v. Asha Rani & Ors.In the case of gratuitous passengers travelling in a goods vehicle, there is no liability of the insurance company at all to pay the compensation. In Padma Sundara Rao & Ors. v. State of T.N. & Ors. (2002) 3 SCC 533. in Herrington v. British Railways Board (1972) 2 WLR 537. It was mentioned in the order that the claimants were very poor man and most of them were daily wage earners. It is the liability of the owner of the tempo i.e. Respondent No.5 (Vinod Kumar) to pay the amount. It is, therefore, directed that the Appellant shall have right to recover the amount of compensation paid to the Respondents (Claimants) in these appeals without having resort to file a civil suit. The Appeals are allowed in above terms.

The Sample result was shown for the legal document summarization of a service law judgement. This contains the main key points that should be present in the final output.

VI. Conclusion

In this paper, An Automatic Legal Document Summarization system is implemented based on the Feature Extraction, using Fuzzy Logic Method. Here 11 top most Feature Extraction methods were used to improve the accuracy of the summary. Single document summarization technique is used in the current system. The system is tested with the input of 150 Legal Judgement documents, in which 50 Documents from the Service Law, 50 Documents from Industry Law, and 50 Documents from Constitutional Law, which is collected from the legal website, The Judgement Information System (http://judis.nic.in/). The result shows that the summary produced by the Fuzzy Logic method gives the complete information about the legal Document in a crisp manner. The quality of the final summary can still be improved by giving a structure to the summary using Rhetorical Roles. In order to understand the structure of a legal Document, it has to segment into coherent paragraphs under some headings using keywords.

Acknowledgement

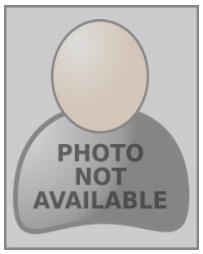
We would like to thank the Legal Expert in the Lexqual as well as Advocate Archana for supporting us.

References

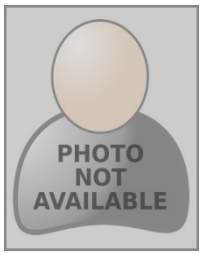
1. A. Archana, C. Sunitha, “An Overview of Document Summarization Techniques”, International Journal on Advanced Computer Theory and Engineering, P: 113-118, 2013.
2. F. Kyoomarsi, H. Khosravi, P.K. Dehkordy, "Optimizing Text Summarization Based on Fuzzy Logic”, IEEE- Computer and Information Science, P: 347-352, 2008.
3. M. Esther Hannah, T.V.Geetha, Saswati Mukherjee, “Automatic Extractive Text Summarization Based on Fuzzy Logic: A Sentence Oriented Approach, Springer-Lecture Notes in Computer Sceince, P: 530-538, 2011.
4. S. Rucha, S.S.Apte, "Improvement of Text Summarization using Fuzzy Logic Based method", IOSR Journal of Computer Engineering, P:5-10, 2012.
5. A.R.Kulkarni, Dr.Apte,”A Domain Specific Automatic Text Summarization using Fuzzy Logic”, International Journal of Computer Engineering & Technology, vol.4,No.4,P:449-461,2013.
6. Maryam Kiabod, Mohammad Naderi Dehkordi, Sayed Mehran Sharafi, “Ä New Effective Criterion to Select Sentences in Extractive Text Summarization”, Journal of Telecommunication, Electronic and Computer Engineering, Vol.4, No.2, P:49-52, 2012.
7. Arman Kiani, Akbarzadeh, “Automatic Text Summarization Using: Hybrid Fuzzy GA-GP”,IEEE International Conference on Fuzzy Systems,2006.

8. M.Saravanan, S.Raman, B.Ravindran, "Improving Legal Document Summarization using Graphical Models.", JURIX 2006, P: 51-60.
9. Ravi Kumar V and K. Raghuveer, Legal Documents Clustering using Latent Dirichlet Allocation, International Journal of Applied Information Systems, 2012, P: 34-37.
10. Claire Grover, Ben Hachey, Ian Hughson, Chris Korycinski: Automatic Summarisation of Legal Documents. ICAIL 2003, P: 243-251.
11. Marie-Francine Moens and Rik De Busser, First steps in building a model for the retrieval of court decisions, International. Journal of Human-Computer Studies, 2002, P: 429–446.
12. Atefeh Farzindar and Guy Lapalme, 'LetSum, an automatic Legal Text Summarizing system' in T. Gordon (ed.), Legal Knowledge and Information Systems, 2004, P: 11-18.
13. S. Santhana Megala, A. Marimuthu, "A Study on Text Summarization Techniques and its Applications", National Conference on Recent Trends and Advances in Information Technology, 2012, P: 5.
14. S. Santhana Megala, A. Marimuthu, "A Comparative Analysis of Legal Text Summarization", International Conference on Design and Applications on Structures, Drives, Communicational and Computing Devices, 2012.
15. S. Santhana Megala, A. Marimuthu, A. Kavitha," Improvised Stemming Algorithm: TWIG", International Journal of Advanced Research in Computer Science and Software Engineering, P: 168-171, 2013.
16. S. Santhana Megala, A. Marimuthu, A. Kavitha, "Enriching Text Summarization using Fuzzy Logics", International Journal of Computer Science and Information Technology, P:863-867,2014S.

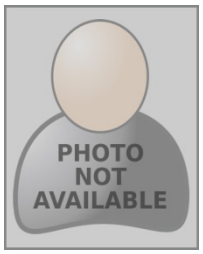
AUTHOR(S) PROFILE



S. Santhana Megala is currently pursuing Ph.D in Computer Science in PRIST University, Thanjavur, Tamil Nadu and working as an Assistant Professor in SNMV College of Arts & Science, Coimbatore, Tamil Nadu, India.



Dr. A. Kavitha is currently working as an Assistant Professor in Dept. of Computer Science, Kongunadu Arts and Science College, Coimbatore, Tamil Nadu, India.



Dr. A. Marimuthu is currently working as an Associate Professor in Dept. of Computer Science, Government College of Arts & Science, Coimbatore, Tamil Nadu, India.

