Multiple Document Summarization Using Text-Based Keyword Extraction

Chapter · February 2016

DOI: 10.1007/978-981-10-0448-3_15

2 authors, including:

Deepak Motwani
ITM University
13 PUBLICATIONS **44** CITATIONS

SEE PROFILE

Multiple Document Summarization Using Text-Based Keyword Extraction

Deepak Motwani and A.S. Saxena

Abstract The main focus of the paper is on the comparison between the proposed methodology keyword-based text extraction using threading and synchronization just like multiple files input as batch processing and previously used technologies for text extraction from research papers. Keyword-based summary is defined as selecting important sentences from actual text. Text summarization is the condensed form of any type of document whether pdf, doc, or txt files but this condensed form should preserve complete information and meaningful text with the help of single input file and multiple input file. It is not an easy task for human being to maintain the summary of large number of documents. Various text summarizations and text extraction techniques are being explained in this paper. Our proposed technique creates the summary by extracting sentences from the original document with the font type and pdf font or keyword extractor.

Keywords Document summarization · Information retrieval · Information extraction · Keyword extraction · Threading · Synchronization batch processing

1 Introduction

In the present era, most of the automated summarization systems create extracts only.

This section explains the various number of problems that we face during retrieval of full text document. There are so many problems associated with full text analysis, but two of them seem to be the most important one these days. First one is the bad quality of search engine mainly the internet search engines and there is a

Deepak Motwani (✉)
Department of CSE, Mewar University, Chittorgarh, Rajasthan, India
e-mail: dmotwani20005@gmail.com

A.S. Saxena
Faculty of Engineering & Technology, Mewar University, Chittorgarh, Rajasthan, India
e-mail: anand.saxena42@gmail.com

M. Pant et al. (eds.), *Proceedings of Fifth International Conference on Soft Computing for Problem Solving*, Advances in Intelligent Systems and Computing 436, DOI 10.1007/978-981-10-0448-3_15

Summarization = Topic Identification + Interpretation + Generation

Fig. 1 General overview of summarization

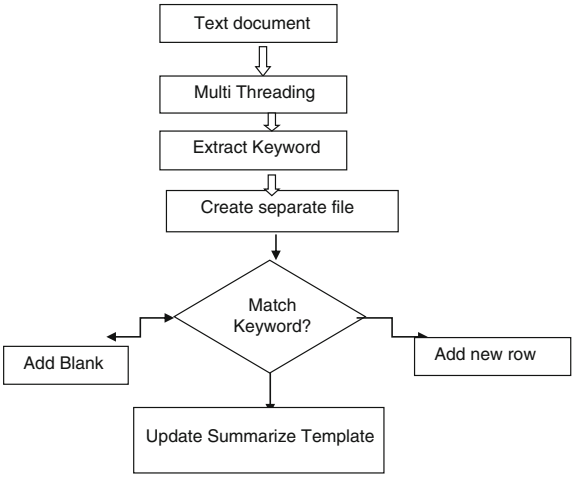

Fig. 2 Flowchart representing algorithm 1

shortage of automatic text categorization tools which are able to do fast assessment of large collection of documents. Many approaches are being used for the extraction in the last 10 years (Figs. 1, 2, 3, 4, 5 and 6).

Abstractive summarization writes the novel sentences which are not seen in the original file. Huge amount of information is present in the form of research publications on the web. Extracted header information is being used for all tasks related to data mining. For the extraction of header information from the research paper, author generates the hybrid approach by combining the given three ParsCit, GROBID, and Mendeley. The proposed methodology is tested on 75 research papers by the author and the precision rate was 95.97 %. Mendeley takes the help of

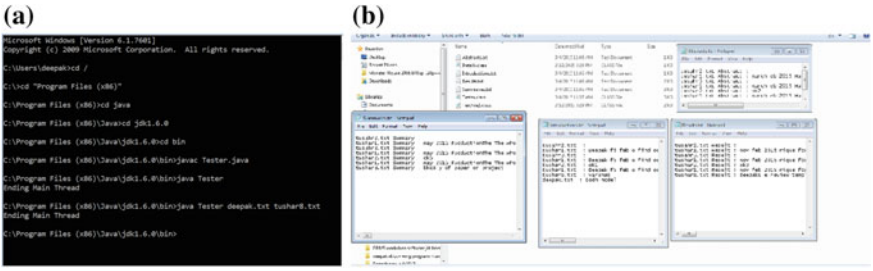

Fig. 3 **a** Command line argument for multiple files and **b** the output in separate files based on keywords

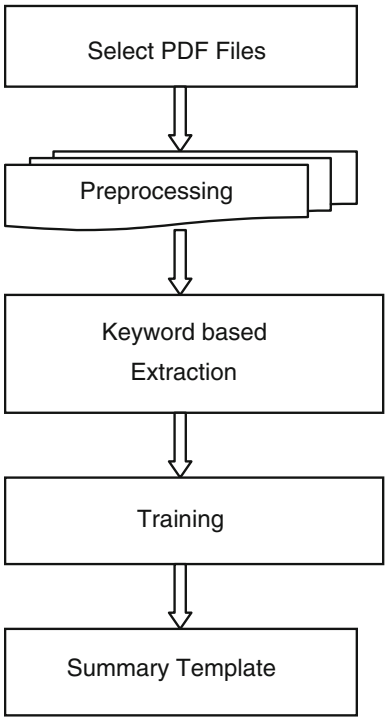

Fig. 4 Flowchart of algorithm 2

SVM and web-based lookup [1] for the citation extraction and embedded metadata from those research papers. Windows-based application is being provided by Mendeley which helps organizing the 111 research papers.

Another advance technique is being used from knowledge base named as Web-based lookup. Trained machine learning algorithm is being used for the authentication of an element in paper, and header information is being extracted. These methods are mainly known as HMM, CRF, and SVM [2]. In this paper, we proposed windows-based application (keyword-based extraction method) for the extraction of title, author, year, and references from research papers so that we can make the summary of research papers.

An artificial intelligence tool which is knowledge-based works in a narrow domain, to give smart decisions, with proper justification. KBS has their own command in the field of artificial intelligence [3].

Example 1 A classic patent analysis scenario

1. Categorization of Task: here one has to define the objective of analysis of task or scope and concept of task.
2. Searching: iteratively filter, seek, and download the linked patents.
3. Segmentation: normalizing the formless and prepared parts, clean, and segment.

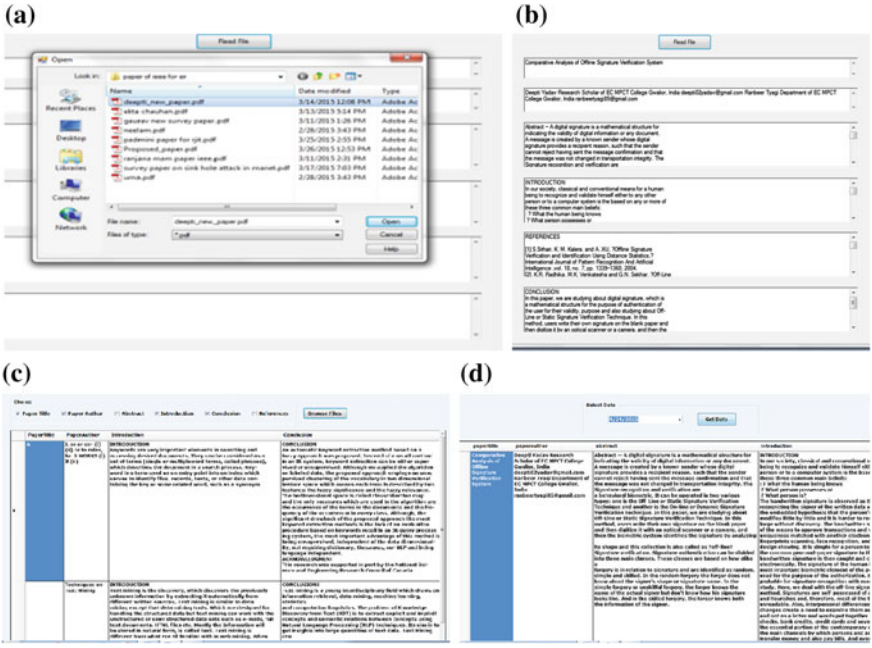

Fig. 5 **a** Represents selection of single/multiple file as an input and **b** represents output template for single file, **c** gives the output template for multiple file and **d** shows the date wise summary report

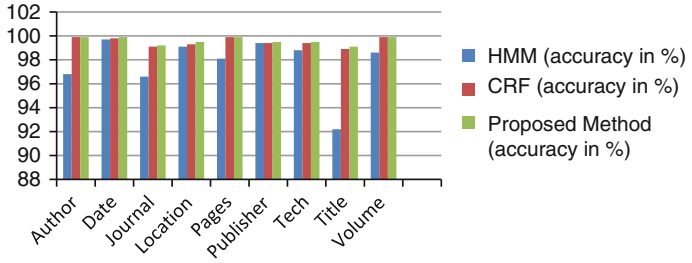

Fig. 6 Graph shows the comparison of accuracy for previous methods and proposed method

4. Abstraction: evaluating the patents content for summarizing their technologies, topics, claims, etc.
5. Clustering: categorizing or gathering of evaluated patent on the basis of extracted attributes.
6. Hallucination: generating the matrices of technology effects.
7. Interpretation: prediction of business trends and relations or technology.

Therefore, on the basis of patent analysis scenario [4] explained above, a text mining methodology dedicated to the analysis of full-text patent. First of all, collect the full patent document which is relevant for analysis purpose.

2 State of Art

In the present time, researchers are again interested in automatic text summarization as they were in fifties. During our literature survey, we came to know that, the paper in which researchers are using the concept of the paper which is published in 1958, the concept used in this paper is that the weight of the sentences of a document is being suggested as a task of high frequency words [5], disregard common words of very high frequency. In 1969, for determining the weight of the sentence "Automatic text summarization system" [6] uses the following three methods:

1. Title Method: this technique evaluates the weight of the sentence; the weight of any sentence is being calculated by summing up the contented words that appear in the title or heading of the text.
2. Cue Method: This technique calculates the weight of sentence by the presence or absence of cue of words.
3. Position technique: the working of this method is done on the basis of some hypothesis that, if a sentence occured in both text and individual paragraph at initial position then there is a high probability of relevance. In 1995, sentence extracting job was done with the help of trainable document summarizer on the basis of number of weighting heuristics [7] following features were used and evaluated:

    a. Sentence Length Cut-O Feature: this represents that those sentences which are containing less than predefined number of words are not considered as the part of abstract.
    b. Fixed-phrase attribute: here sentences are having few cue of words and phrases are also taken.
    c. Paragraph attribute: works as location method attribute [8]
    d. Thematic Word attribute: the most recurrent words are named as thematic words. Thematic words are having functions named as scores of words.

The ANES text extraction system [9], in 1995, is a system that performs automatic, domain-independent compression of news data. The process of summary generation has three major constituents:

1. Sentence weighting, 2. Corpus analysis, and 3. Selection of sentences

Hidden markov model (HMM) [5]: is the powerful mathematical statistical tool for document retrieval. IR can employ the subparts of the document like building links [10] and/or clusters between index terms and so on. Although this is not an issue in any of the above mentioned abstracting systems, it seems to be worth of consideration when building such systems (Table 1).

Table 1 Literature survey

Year	Author	Title	Methodology	Objective
1997	Thorsten Joachims	Text categorization with supports vector machines: learning with many relevant features [11]	Support vector machines (SVMs)	Text categorization
2004	Masoud Makrehchi	Fuzzy set approach to extract keywords [12]	Fuzzy set approach	Keyword extraction
2009	Rasim Alguliev	Evolutionary algorithm for extracting text summarization [13]	Discrete differential equation	Optimization of objective function
2012	Shu-Hsien Liao	Data mining techniques and applications [14]	Data mining techniques	Ability to continually change and provide new understanding by DMT
2012	Ozair Saleem	Information extraction from research papers by data integration and data validation from multiple header extraction sources [15]	(HybridMethod) GROBID, ParsCit, and Mendeley	To achieve accurate header extraction
2012	Hao Lul	Research on intelligent scientific research collaboration platform and taking journal intelligence system [16]	Intelligent journal prototype system	Prototype system can be changed to reliable, available, and effective scientific research collaboration platform
2012	Yogan Jaya Kumar	Automatic multi-document summarization approaches [17]	feature-based method, cluster-based method, graph-based method and knowledge-based method	To generate a better summary this is well suited for an informative type summary generation
2013	LI Yan-min	Applying information retrieval technology in analyzing the journals [18]	Variable of index term operators	CNKI's database searching
2013	Joeran Beel	Docear's PDF inspector: title extraction from PDF files [19]	Docear's PDF inspector	To extract titles from academic PDF files by applying a simple heuristic: the largest text on the first page of a PDF is assumed to be the title
2014	Xianfeng Yang and Liming Lian	A new data mining algorithm based on map reduce and Hadoop [20]	Map-reduce programming model and Hadoop, Newman algorithm	Discovering hidden information from large databases with respect to scalability

3 Problem Statement

In full text analysis, two problems seem to be the most important these days. They are firstly the bad quality of search engines mainly internet search engines and second one is the shortage of text categorization tools which permit the fast assessment for large number of documents. To deal with information explosion and to text, categorization will be the best technique. Text categorization would be much easier to deal with "information explosion" and assimilate all data that is going to flood us, if we are able to find out the main subject from the document and then arrange it into some sequential manner or say in some structured form, preferably hierarchical. The most traditional technique used for this problem is building of handcrafted index. Building a handcrafted index would be the traditional approach to this problem; in fact these indexes are of extensive use amid of the juridical communities and Internet. Unluckily, they merely cannot deal with number of new documents formed everyday. Due to the growing increment of the availability of large amount of information it seems incomplete and index creator is able to classify and analyze it. So, automatic text categorization is strongly required here. In this paper, we will build an extraction keyword-based text summarizer based on windows application that will provide a multiple text file using command line, create a summary of the original text in different files; run tester Java file. To build an automated text summarizer, assign the variable keywords to the set of significant words in the document.

4 Proposed Methodology

In this paper, we proposed keyword-based text extraction technique using threading and synchronization just like multiple files input as batch processing. Whole process is explained with the help of flowcharts and algorithms.

Experiment 1 Text files as an input (single document or multiple documents)

Algorithm 1 For .txt file

1. The Main module: Calls the thread (multithread) module to produce the summary of a text in different file using extract keyword. Main module (input f1, f2, f3, …, fn)
2. The Generator module contains the following methods: set Keywords generate summary: prints out the most significant sentences in the document. Create module (set keyword, limit summary)
3. Create Introduction, abstract, conclusion, and separate summary file.

Switch module
If == match keyword (for multiple files)
Then
Create separate abstract file
Create separate Introduction file
Create separate conclusion file

4. Input files like tushar.txt applied as command line argument
5. Separate (keyword, font, and style) text file created on the basis of keyword matching from abstract, introduction, result, and summary after program execution using thread and synchronization
6. Output summary file public class details which contains String intro, abstracts, result, and summary; and make it public class tester implements runnable.

Create separate summary file

Experiment 2 pdf or doc files as input (single document or multiple documents)

Algorithm 2

The main module: reader application and class PDF Form: Form and doc or word document form
Main module(input file type doc, pdf, other)

The generator module contains the following methods: Choices for keyword: paper title, paper author, abstract, introduction, conclusion, and references. Read (keyword1, keyword2, keyword3, …)

If word == match
Match (file1, file2, file3.)
Else
No match ("word not match")
Then
pattern (title, abstract, introduction, reference, summary…..)
If pattern == title (Title extract: using font-size)
{
abstract paragraph fetch and store in data base
}
Else if pattern == reference(Match keyword and extract contents stored into specified template)
{
References fetch and store in data base
}
Else if pattern = introduction
{
Introduction paragraph fetch and store in data base
}
5. Display in template and view and exit

5 Experiment Results and Analysis

We have taken the datasets of 1000 pdf, doc, txt files for the extraction purpose and coding is done in Java, .Net, and C# programming. We compare our result with the result of other three methods used in previous papers. Out of these 1000 pdfs, SciPlore Xtract cannot extract data from 307 pdfs. These 307 pdfs consist of scanned images and OCR is being applied on them so SciPlore Xtract found difficulty in extraction. So, SciPlore Xtract works with 693 pdfs for the analysis. While comparing the result, we see that from 693 pdfs titles could not be extracted for 54 pdfs by using SciPlore or CiteSeer's SVM. All three approaches are able to identify the titles of only 160 pdfs. In final analysis, we observe that SciPlore Xtract can extract the titles of 540 pdfs correctly (77.9 %). CiteSeer's SVM can identify 481 titles correctly. SciPlore Xtract takes 8:19 min for extracting the titles and SVM needed 57 and our proposed methodology can extract titles from 640 pdfs correctly, 1.4 % is the error in the extraction and time taken for it is very less around 6:40 min. So, our results show the efficiency of the proposed methodology. Table 2 shows the result comparison of all the four methods for extracting the title and Table 3 represents the accuracy comparison chart extraction results for paper references from previous method and proposed system and figure shows the graphical representation of both the tables result. Overall accuracy for the results of HMM is 75 %, CRF is 99.5 %, and keyword-based extraction method (proposed method) 99.7 %.

Table 2 Title Extraction of 693 pdfs, (doc, txt, pdf: three format of file added in our work)

Methods	Correct		Slight errors		Total	
SciPlore Xtract	528	76.2 %	12	1.7 %	540	77.9 %
CiteSeer SVM + pdftotext	406	58.6 %	75	10.8 %	481	69.4 %
CiteSeer SVM + PDFBox	370	53.4 %	78	11.3 %	448	64.6 %
Keyword based extraction (proposed method)	630	90.9 %	10	1.4 %	640	92.3 %

Table 3 Extraction results for paper references from previous method and proposed system

Keywords	HMM (accuracy in %)	CRF (accuracy in %)	Proposed method (accuracy in %)
Author	96.8	99.9	99.9
Date	99.7	99.8	99.9
Journal	96.6	99.1	99.2
Location	99.1	99.3	99.5
Pages	98.1	99.9	99.9
Publisher	99.4	99.4	99.5
Tech	98.8	99.4	99.5
Title	92.2	98.9	99.1
Volume	98.6	99.9	99.9
Average accuracy	75	99.5	99.7

6 Conclusion and Future Scope

This paper made a clear and a simple overview of working of text extraction from pdf document in step by step process. We know that many text extraction systems are available in the market but researchers are working in this area to improve the efficiency because till now we find difficulty in extracting the text from the document containing tables, images, and so on. So the expectation of researchers is to extract the text from complex document very smoothly. Our method is able to work on these parameters and the accuracy is much better as compared to previous results. The accuracy percentage for the extraction for paper references from pdfs is approximately 99.7 % and for title extraction its 92.3 % and time taken for the extraction is very less. In future, we try to work using some hybrid techniques.

References

1. Mendeley is a desktop and web program for managing and sharing research papers, discovering research data and collaborating online
2. Accurate Information Extraction from Research Papers using Conditional Random Fields
3. Lin, C.-J., Lin, Y.-I.: Text mining techniques for patent analysis. Int. J. Inf. Proc. Manag., ACM, USA, **43**, 1216–1247 (2007)
4. Tu, Y.-N., Seng, J.-L.: Research intelligence involving information retrieval—an example of conferences and journals. Int. J. Expert Syst. Appl. 12151–12166 (2009)
5. Luhn, H.P.: The automatic creation of literature abstracts. Int. J. IBM J. Res. Dev., ACM, USA, vol. 2, pp. 159–165, 1958.
6. Edmundson, H.P.: New methods in automatic extracting. J. ACM, USA **16**, 264–285 (1969)
7. Kupiec, J., Pedersen, J., Chen, F.: A trainable document summarizer. In: Proceedings of the 18th ACMSIGIR Conference on Research and Development in Information Retrieval, USA, pp. 68–73 (1995)
8. Mittendorf, E., Schauble, P.: Document and passage retrieval based on hidden markov models. In: Proceedings of the 17th ACM-SIGIR Conference on Research and Development in Information Retrieval, New York, pp. 318–327 (1994)
9. Brandow, R., Mitze, K., Rau, L.F.: Automatic condensation of electronic publications by sentence selection. In: International Journal on Information Processing and Management, ACM, USA, vol. 31, pp. 675–685 (1995)
10. Bookstein, A., Klein S.T., Raita, T.: Detecting content-bearing words by serial clustering. In: Proceedings of the 18th ACM-SIGIR Conference on Research and Development in Information Technology, New York, pp. 319–327 (1995)
11. Joachims, T.: Text categorization with support vector machines: learning with many relevant features. In: Proceedings of European Conference on Machine Learning, ACM, London, pp. 137–142 (1998)
12. Makrehchi, M., Kamel, M.: A fuzzy set approach to extracting keywords from abstracts. IEEE Int. Conf. Fuzzy Inf. **2**, 528–532 (2004)
13. Alguliev, R., Aliguliyev, R.: Evolutionary algorithm for extractive text summarization. Int. J. Intell. Inf. Manag. **1** (2), 128–138 (2009).
14. Liao, S.-H., Chu, P.-H., Hsiao, P.-Y.: Data mining techniques and applications– A decade review from 2000 to 2011. J. Expert Syst. Appl., Elsevier **39**, 11303–11311 (2012)

15. Saleem, O., Latif, S.: Information extraction from research papers by data integration and data validation from multiple header extraction sources. In: World Congress on Engineering and Computer Science (WCECS), San Francisco, USA (2012)
16. Lu, H., Zheng, X., Sun, X., Zhang, N.: Research on intelligent scientific research collaboration platform and taking journal intelligence system as example. In: International Conference on Service Operations and Logistics, and Informatics (SOLI), IEEE, Suzhou, pp. 138–143 (2012)
17. Kumar, Y.J., Salim, N.: Automatic multi document summarization approaches. Int. J. Comput. Sci.
18. Xie, W.-L., Li, Y.-M., Zhang, Y.: Applying information retrieval technology in analyzing the journals. In: Fourth International Conference on Emerging Intelligent Data and Web Technologies (EIDWT), Xi'an, pp. 88–94 (2013)
19. Beel, J., Langer, S., Genzmehr, M., Müller, C.: Docear's PDF inspector: title extraction from PDF files. In: Proceedings of 13th ACM/IEEE-CS joint Conference on Digital Libraries, ACM, USA, pp. 443–444 (2013)
20. Yang, X., Lian, L.: A new data mining algorithm based on map reduce and Hadoop. Int. J. Signal Process. Image Process. Pattern Recogn. **7**, 131–142 (2014)