Detecting the Originality of Extended Research Articles Using Similarity Techniques – A Proposal

Shanmugasundaram Hariharan
B. S. Abdur Rahman University, Chennai, Tamilnadu, India

***Abstract*— Research articles with respect to increased growth of research all round the world, personal interest of academicians, industrialists or other personals has paved way to huge knowledge repositories. The outcome of knowledge base provided by such a group is both advantageous and vice versa. In sense we mean that if the knowledge were used properly, it leads to innovation of new ideas or further enhancements, while on the other side when such knowledge is used repetitively in a reformulated fashion it would result in plagiarism of research articles pulling out the quality of research. This article proposes a method to detect the originality of research articles published as an extension of the previous works using similarity technique. We have proposed a method to summarize the scientific papers, so as to reduce the time involved in reading a document, there by presenting the originality of the research artciles. From the preliminary study carried out using the corpus collected, the study seems to be promising. We focus on to implement the proposed system and provide a comparison of the methods discussed in this paper analyzing the necessary parameters.**

***Index Terms*— Similarity measures, Cosine metric, scientific articles, summarization, extended research papers**

I. INTRODUCTION

Text summarization is a technique of automatically creating summary from one or more texts. Initial interest for automatic summarization started in 1960's in American research libraries, where a large amount of scientific papers and books were to be digitally stored and made searchable. Research on automatic summarizing, taken as including extracting, abstracting, etc., has a long history with an early burst of effort in the sixties following Luhn's pioneering work [4], followed by marked growth of activity since the mid eighties and especially very recently[5]. Research on summarization has attained its roots long back in 1950's-1960's and has become a steady subject of interest among research community [4, 7].

Text summarization has several branches or dimensions [13].History of summarization can be traced through several approaches like surface –level approaches in 1950's, entity – level approaches in 1960's, extensive entity level approaches in 1970's, 1990's with the renaissance of all three approaches [14]. In addition we are seeing the emergence of new areas such as multi-document summarization, multilingual summarization and multimedia summarization [13].

Due to strenuous efforts and improvements over the significant years [15], summarization has been a major challenge and has drifted attention among the research community. Rapid advancement of Internet technologies, documents available on the web were digitized and made available online. Researchers regularly browse through literature papers to update their existing knowledge or for use in their research work. The challenge lies in summarizing the research articles. This paper mainly focuses on methods to detect the similarity among research articles. The proposed idea well suits for identifying the originality of papers submitted for conference submissions, journals and more specifically extended research articles that are submitted for publication in more than one system. The proposed system leads to the following.

- To prevent malpractice of repeating the same article in repetitive forms.
- Helps the researchers to think in a broader sense.
- Improves the quality of the paper.
- Helps the academicians to enhance the work in better way.

The paper is organized as follows: Section II discusses the related research carried out. In Section III, proposed system with sample corpus, modules involved in the proposed system are explained. Finally Section IV gives the conclusions.

II. RELATED WORK

Simone Teufel et al., [1] proposed a system for scientific articles that concentrates the statements which have rhetorical status for summarization and highlighted summaries. Several experiments were conducted for substantial corpus of conference articles with human judgments of rhetorical status and sentence relevance. Extraction of unseen articles content and classification into fixed set of seven numbers of rhetorical categories is done which is viewed further as a single-document summary.

Dain Kaplan et al., [2] designed an automatic method based on co-reference-chains for extracting citations from research papers. The authors considered the span of text as "c sites" which describes the work being cited. The system is designed

Shanmuasundaram Hariharan is with the Department of Information Technology, B.S.Abdur Rahman University, Chennai, Tamilnadu, India.(Phone :04422751347, Mobile: +91-9884204036, E-mail : mailtos.hariharan@gmail.com). He is working as Assistant Professor and currently pursuing his doctoral programme in the area of Information Retrieval.

to extract all c-sites that refer the target paper is aggregated later to form summary. It is different from traditional summarization technique based on fact the summary generated contains multiple points-of-view. The authors conducted several surveys in relation to parsing and extraction on several pre-existing components.

Vahed Qazvinian et al., [3] developed a model that summarizes a single article that can be summarized further based on the topic. The author found that this model breaks the difficulty for researchers to explore the vast amount of scientific literature in each field of study. Authors have used some effective clustering approaches for study of citation summary network based on the views given to the articles by others.

Balage Filho et al., [5] presented a paper with experiments on scientific text summarization. The authors adopted the simple extractive summarizers that generates a small version of main part from the given complete source text. It is also found from their investigation that by considering the specificity of the text genre, the results can be improved and made better. The authors have validated summarization process by taking only text structures.

Maher Jaoua et al., [6] proposed a system which creates indicative summaries for scientific papers that differs from conventional methods. The authors extract summaries in two steps. First step produce a population of extracts followed by second step that classifies and selects the best one based on some global criteria that are defined for whole extract than sentences. The authors have deployed a summarization system for French language called "ExtraGen" is developed as a prototype that performs the generation and classification mechanism by implementing a genetic algorithm.

Stephen Wan et al., [8] developed a new research tool called Citation-Sensitive In-Browser Summarizer (CSIBS), to speed up the update on research article based on user requirement browsing task. Building such comprehensive summary helps the readers to explore the cited article and determine whether time is to be spent and thereby alleviate overload. The authors retrieved the sentences from cited document by exploiting citation context by bringing together the meta-data and a citation-sensitive. The authors found that the relevancy judgment task is facilitated by CSIBS, thereby the users' self reported confidence in making such judgments is increased.

II. EXPERIMENTAL SETUP

This section briefs the proposed system to detect the originality in extended research articles. Fig 1 gives the proposed architectural system, with each module discussed in several subsections. The input source article is mostly in pdf or html format. We convert the documents in either form to text for further processing. Table I presents the details of the samples investigated for the proposed system.

The corpuses were collected from the papers submitted by different authors to International Conference and journals, which we have surveyed personally for project and research activities (This proposed system is currently under implementation). A sample set S1 is shown in Figures 2A, 2B, 2C for illustration. The proposed system would provide a solution to point out, which of the articles is better or worth reading, which document has been plagiarized etc.

TABLE I: DATA CORPUS AND STATISTICS

Set ID	Title	Year of publication	Authors	Publisher
S1	Centroid Based summarization	2000	Dragomir Radev	ACL
	Centroid Based summarization of multiple documents	2004	Dragomir Radev	Elsevier
	Centroid Based summarization of multiple documents Using timestamps	2007	Nedunchelian	IEEE
S2	A Bottom up approach for sentence ordering in multi document summarization	2007	Danuksha Bollegola	IEEE
	A Bottom up approach for sentence ordering in multi document summarization	2009	Danuksha Bollegola	Elsevier

A. Pre-processing Of Documents:

Preprocessing is most significant task in data mining tasks, information retrieval and several other tasks. In this phase the documents were pre-processed by the following steps:

a. Converting all uppercase letters to lowercase.
b. Removing unwanted words (stop words)
c. Stemming the terms occurring in the document.

To measure the content similarity (discussed in Section B), we remove the stop words from the text document followed by stemming of samples. We adopt the same process whenever similarity is measured.

Stop words are ordinary or unusual words which occur in the document, which don't have significant meaning (e.g. connector words, conjunctions, single letter words). From a corpus of database, we eliminate such unwanted words [10]. We also eliminate special symbols that do not have significant part in text processing (e.g. ",", /,-, etc, in general symbols other than characters and numbers).

Truncation, also called stemming, is a technique that allows us to search for various word endings and spellings simultaneously. Stemming algorithms [11] are used in many types of language processing and text analysis systems, and are also widely used in summarization, information retrieval and database search systems.

A stemmer is a program determines a stem form of a given word. Terms with a common stem will usually have similar

meanings. An example is shown below, that gives to a common stem:

ACUSES, ACUSER, ACUSED, ACUSERS = ACCUS

The suffix stripping process will reduce the total number of terms in the IR system, and hence reduce the size and complexity of the data in the system, which is advantageous. We have adopted suffix stripping that would help us better in term frequency calculations during sentence scoring. Each stem term and its equivalents are stored in database and retrieved whenever comparison is required [11].

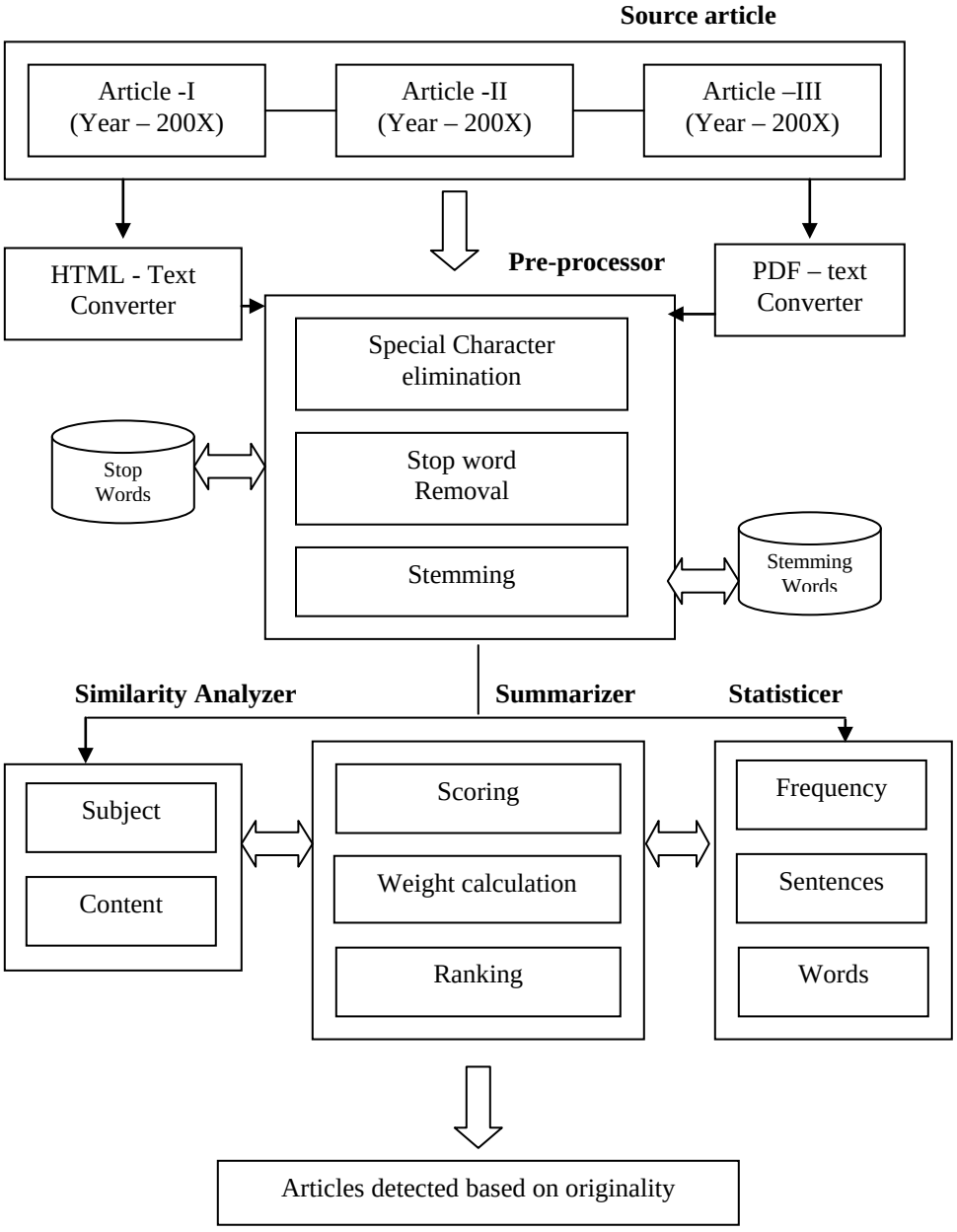


Fig 1: Proposed system Architecture

B. Similarity analysis:

The system designed to detect similarity among text documents calculates content similarity among specified documents. The similarity is estimated between the articles using cosine metric. Though there exist several choices like dice, Jaccard, Hellinger, we adopt cosine measure, which is quiet popular and yields better results [9,12] given by expression (1).

$$Cosine(ti,tj) = \sum_{h=1}^{k} t_{ih} t_{jh} / \sqrt{\sum_{h=1}^{k} t_{ih}^2 \sum_{h=1}^{k} t_{jh}^2} \quad \text{---(1)}$$

A document is treated as a vector, with t_{ih} representing the first vector (each term being normalized using the total number of words in the document) and t_{jh} corresponds to the second vector, k: is the number of terms in the document. Statisticer acts as agent in finding out the number of sentences, word count and frequency of words. Based on the term frequency of both documents, cosine metric obtains a value in the interval of 0-1. If both documents are exactly same they have a value of 1, 0 otherwise.

Consider an example to illustrate the calculations for measuring the similarity. If S1 and S2 denotes two different

sentences correspondingly, then cosine measure is calculated as shown below.

S1: Heavy earth quake affects Indonesia.

S2: Indonesia rocked at 6.5 ritcher scale.

All the terms in each sentence have term frequency of 1. Hence cosine value = (1*1)/sqrt (5*6) = 1/ 2.23* 2.44 = 0.18. A suitable threshold can be fixed to identify the similarity among extended and original articles.

This section also discusses the issue of measuring the similarity among documents. Number of papers available online, lack of commercially available tools to detect the plagiarism effectively has made us to focus on such an issue. Our focus lies on measuring the similarity of documents under the following categories.

- *Similarity measured as whole (Category – I)*
  - measures the similarity by taking the entire content
- *Similarity measured under each subject/criteria(Category – II)*
  - Each technical paper is structured in different ways. However the paper has titles like "Abstract", "Conclusion/Future Work", "Introduction", "Results", "Experimental Section"
- *Similarity measured from summaries of each articles(Category – III)*
  - Generates summary for each criteria
  - Measures the ratio of similarity under each category

Through these categories, we would set penalties for sentence scoring. Consider an scenario to illustrate the situations of each the categories discussed above. Category – I when it measures the relevancy among the entire contents, may not reflect the originality of the articles. The reason behind this that users who write research articles abstract the contents from the original source. Hence we may go for Category II and measure the importance at each level, which would lead to better conclusions. This would allow the researchers or academicians who are interested in reading out the literature to skip the paper immediately or to read anyone (which is worth while) without wasting his time. Category – III would be more useful in upcoming years, as literature papers have been increasing year by year. Hence determining the similarity among research articles based on summary of each individual article may be deemed useful, provided the quality of generated summary is good.

C. *Summarizer:*

The algorithm designed for summarizing scientific articles consists of the following steps:

a. Scoring each sentence in document.
b. Weight calculation for each scored sentence.
c. Ranking the results based on the user requirements.

Sentences within each document are ranked depending on the sentence weights using term frequency approach [8]. For generating weight for each sentence we adopted extraction method, where we measure the frequency of terms in each document with special weights considered for each special category like (bold, italic, words matching title or subtitle etc...). Finally summary is generated depending on the score each sentences obtain.

IV. CONCLUSIONS

The paper has focused on the issue of research articles published in different publications. We have identified a solution to detect the similarity of the research articles, so as the proposed system might provide assistance in finding out the originality of the content. The system also serves as a platform to detect plagiarism, especially for papers submitted for conferences. The system can also well be adopted to measure the relativeness among articles of any nature depending on the users choice by modifying the threshold. The paper has not focused on the implementation part, as it is currently implemented. We also focus to identify main theme of each articles based on the summaries of each articles.

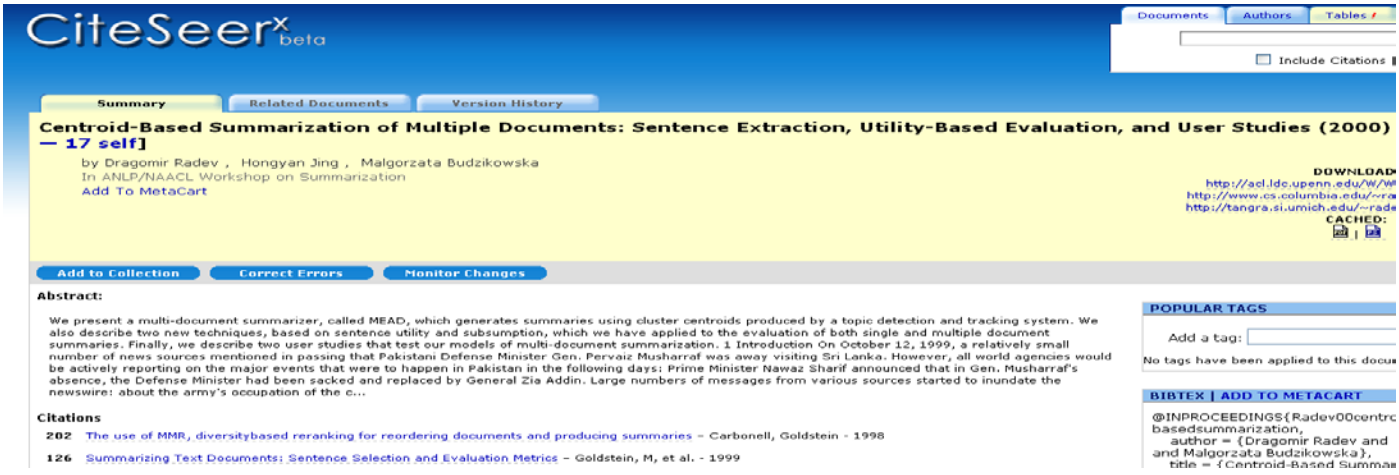


Fig 2A: Paper published in the Year 2000 – In Proceedings of ANLP/NAACL Workshop

Centroid-based summarization of multiple documents

Dragomir R. Radev[a], [b], Hongyan Jing[b], Małgorzata Styś[b] and Daniel Tam[b]

[a] University of Michigan, Ann Arbor, MI 48109, USA

[b] IBM T.J. Watson Research Center, Yorktown Heights, NY 10598, USA

Received 6 January 2003; accepted 24 October 2003. Available online 13 December 2003.

Abstract

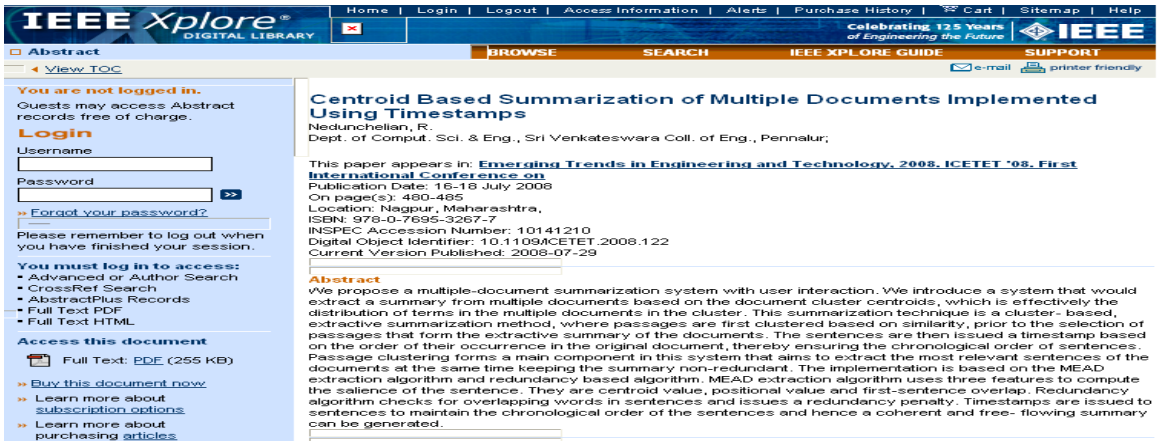
We present a multi-document summarizer, MEAD, which generates summaries using cluster centroids produced by a topic detection and tracking system. We describe two new techniques, a centroid-based summarizer, and an evaluation scheme based on sentence utility and subsumption. We have applied this evaluation to both single and multiple document summaries. Finally, we describe two user studies that test our models of multi-document summarization.

Author Keywords: Multi-document summarization; Centroid-based summarization; Cluster-based relative utility; Cross-sentence informational subsumption

Article Outline

1. Introduction
1.1. Topic detection and multi-document summarization
2. Informational content of sentences
2.1. Cluster-based relative utility (CBRU)
2.2. Cross-sentence informational subsumption (CSIS)

Fig 2B: Paper published in the year 2004 by Elsevier



Centroid Based Summarization of Multiple Documents Implemented Using Timestamps

Nedunchelian, R.
Dept. of Comput. Sci. & Eng., Sri Venkateswara Coll. of Eng., Pennalur;

This paper appears in: Emerging Trends in Engineering and Technology, 2008. ICETET '08. First International Conference on
Publication Date: 16-18 July 2008
On page(s): 480-485
Location: Nagpur, Maharashtra,
ISBN: 978-0-7695-3267-7
INSPEC Accession Number: 10141210
Digital Object Identifier: 10.1109/ICETET.2008.122
Current Version Published: 2008-07-29

Abstract

We propose a multiple-document summarization system with user interaction. We introduce a system that would extract a summary from multiple documents based on the document cluster centroids, which is effectively the distribution of terms in the multiple documents in the cluster. This summarization technique is a cluster- based, extractive summarization method, where passages are first clustered based on similarity, prior to the selection of passages that form the extractive summary of the documents. The sentences are then issued a timestamp based on the order of their occurrence in the original document, thereby ensuring the chronological order of sentences. Passage clustering forms a main component in this system that aims to extract the most relevant sentences of the documents at the same time keeping the summary non-redundant. The implementation is based on the MEAD extraction algorithm and redundancy based algorithm. MEAD extraction algorithm uses three features to compute the salience of the sentence. They are centroid value, positional value and first-sentence overlap. Redundancy algorithm checks for overlapping words in sentences and issues a redundancy penalty. Timestamps are issued to sentences to maintain the chronological order of the sentences and hence a coherent and free- flowing summary can be generated.

Fig 2C: Paper published in the year 2008 by IEEE

REFERENCES

[1] Simone Teufel and Marc Moens, "Summarizing Scientific Articles:Experiments with Relevance and Rhetorical Status", In: Association for Computational Linguistics,Vol .28,No. 4,pp 409-445, 2002.

[2] Dain Kaplan, Ryu Iida and Takenobu Tokunaga, "Automatic Extraction of Citation Contexts for Research Paper Summarization: A Coreference-chain based Approach", Proceedings of the 2009 Workshop on Text and Citation Analysis for Scholarly Digital Libraries, ACL-IJCNLP 2009, pp. 88–95, Suntec, Singapore, 2009.

[3] Vahed Qazvinian and Dragomir R. Radev, "Scientific Paper Summarization Using Citation Summary Networks", Proceedings of the 22nd International Conference on Computational Linguistics (Coling 2008), pp. 689–696, Manchester, 2008.

[4] Luhn, H. P,"The Automatic Creation of Literature Abstracts". IBM Journal of Research Development, 2(2):159–165,1958.

[5] Balage Filho, P.P. Salgueiro Pardo, T.A. and das Gracas Volpe Nunes, M, "Summarizing Scientific Texts: Experiments with Extractive Summarizers" ,In Proceedings of IEEE,Intelligent Systems Design and Applications, 2007. ISDA 2007. Seventh International Conference , pp.520-524, 2007.

[6] Maher Jaoua and Abdelmajid Ben Hamadou, "Automatic Text Summarization of Scientific Articles Based on Classification of Extract's Population" , In proceedings of Computational linguistics and intellignet text processing, Lecture Notes in Computer Science, Springer Berlin / Heidelberg, Volume 2588/2008, pp. 363-377, 2008.

[7] H.P. Edmundson, " New Methods in Automatic Extracting ", Journal of the ACM, Vol .16, no 2 ,pp.264-285, 1969.

[8] Shanmugasundaram Hariharan and Rengarmanujam Srinivasan, "Investigations in Single document Summarization by Extraction Method", In Proceedings of IEEE International Conference on Computing, Communication and Networking (ICCCN'08), pp.1-5, 2008.

[9] Shanmugasundaram Hariharan and Rengaramanujam Srinivasan (2008b), "A Comparison of Similarity Measures for Text Documents", Journal of Information & Knowledge Management, Vol. 7, No. 1, pp. 1–8.

[10] http://ir.dcs.gla.ac.uk/resources/linguistic_utils/stop_words(Last accessed on 2 December 2009)

[11] M.F. Porter, "An algorithm for suffix stripping", *Program*, 14(3) pp 130−137, 1980.

[12] Michael W.Berry, Murray Brown , "Lecture notes in Data Mining", World Scientific Publishing., 2006.

[13] Mani.I., and M.T Maybury , "Advances in Automatic Summarization.", MIT Press , Cambridege, MA., 1999.

[14] Karen Sparck-Jones (2007).Automatic Summarizing: The state of art. Information Processing and Management ,Issue :43 , pp:1449-1481, 2007.

[15] Karel Jezek and Josef Steinberger, "Automatic summarization (The state of Art 2007 and new challenges)", Vaclav Snasel, Znalosti, pp.1-12.,2008.