Global Journals LATEX JournalKaleidoscope™

Artificial Intelligence formulated this projection for compatibility purposes from the original article published at Global Journals. However, this technology is currently in beta. *Therefore, kindly ignore odd layouts, missed formulae, text, tables, or figures.*

Automatic Multiple Document Text Summarization Using Wordnet and Agility Tool

naresh kumar[1]

[1] GGSIPU

Received: 16 December 2013 Accepted: 3 January 2014 Published: 15 January 2014

Abstract

The number of web pages on the World Wide Web is increasing very rapidly. Consequently, search engines like Google, AltaVista, Bing etc. provides a long list of URLs to the end user. So, it becomes very difficult to review and analyze each web page manually. That?s why automatic text sumarization is used to summarize the source text into its shorter version by preserving its information content and overall meaning. This paper proposes an automatic multiple documents text summarization technique called AMDTSWA, which allows the end user to select multiple URLs to generate their summarized results in parallel. AMDTSWA makes the use of concept based segmentation, HTML DOM tree and concept blocks formation. Similarities of contents are determined by calculating the sentence score and useful information is extracted for generating a comparative summary. The proposed approach is implemented by using ASP.Net and gives good results.

Index terms— document text summarization, web page, similarity, summarizer, www, DOM tree, word net, agilitytool.

1 Introduction

he number of documents and users on the World Wide Web (WWW) is increasing with a very high speed. This increases the size of any repository of search system to a very large extent. The search system like Google provides a large number of URLs corresponding to the search keywords. The results retuned by the Search Engine (SE) contain a small description of the text also. But, such snippets are limited to at most three lines of text. Moreover, these lines are the initial line of the document which may or may not provide some meaningful information to the end user. That's why automatic text summarization (ATS) techniques are used [1]. This helps the end user in understanding the main ideas of documents quickly [2] [3]. The task of summarization is classified into two types **??**4] i.e. single document text summarization and multi-document text summarization. But the study of [5] showed that, after 2002 the use of single document summarization was almost dropped. Now multidocument text ummarization techniques are in use. In this technique several issues like reducing each document up to some extent, incorporating major significant thoughts and suggestions, ordering of the sentences coming from different sources by keeping the logical and grammatical structure in proper format [6]. This paper presents AMDTSWA to address these issues. Rest of paper is organized as: section 2 describes the related work, section 3 and section 4 describes the problem formulation and proposed approach respectively. Section 5 and section 6 explain experimental setup and achieved results correspondingly. Section 7 concludes the paper.

2 II.

3 Related Work

Query sensitive text summarization technique that can provide the summary of single or multiple web pages was purposed in [7]. There user could select a set of links from the search engine results and then text summarizer returned the summary of selected links. Concept based segmentation technique utilized the Document Object Model (DOM) tree to analyze the contents of the web page. The leaf node of this tree was called micro block and adjacent micro block were merged to form a topic block. Each of these sentences were labeled by using ASSERT software. Topic blocks containing information about similar concept word were merged to form a concept block. The results were arranged in descending order of sentence similarity score. The top scoring sentences were extracted and their corresponding web pages were arranged in hierarchical structure. The experimental results proved to be superior in terms of control over the results, quick decision making and reduction of time complexity during processing. But nothing was done on tabular data.

Multiple document text summarization technique for improving the effectiveness of retrieval and accessibility of e-learning was purposed in **??**8]. The original document was partitioned into range block and then transformed into a hierarchical tree structure. The range block was represented by nodes of the tree. Then the number of sentences according to the comparison ratio was extracted and some significance score was assigned to them. In traditional summarization techniques; the importance of any sentence was indicated by its location. But today, the textual information like news inside a node was considered equally important regardless of its location inside the node. Therefore, the location feature was not considered during hierarchical summarization of the tree structure. The results of proposed work were tested using t-test and found more superior than the existing system of summarization. T multi document text summarization. CPSL technique was combination of MEAD and Sim With First feature. The similarity score of each sentence with respect to first sentence was computed. Then the highest score was chosen as the most similar sentence. At last, the cosine similarity between a sentence at specific position and the first sentence in the document was calculated. Then MEAD decides which sentence to include in the summary on the basis of sentence's score. The LESM technique was the combination of LEAD and CPSL. At initial level summery of text was generated according to LEAD and CPSL techniques. Then common sentences from the summaries of both summarizers were chosen. The last sentence of a document was considered for concluding the document. At the end authors claimed that for single and multi document text summarization CPSL can provide better results than MEAD. Furthermore, LESM can provide better results for short summaries, but also agreed on better quality of CPSL.

A technique for multi-document text sumarization using mutual reinforcement and relevance propagation models was proposed in [10]. It provides the addition of features to sentences with existing query and Reinforcement After Relevance Propagation (RARP). The architecture of RARP consists of three steps i.e. Pre-processing, sentence score calculation based on feature profile and sentence ranking by reinforcement. Pre processing step consi-dered .txt, .pdf, .rtf, .doc, .html etc. and query as input. Sentence score was calculated using term feature formula. Sentence ranking by RARP and sentence extraction was achieved by using manifold ranking based algorithm. After ranking of sentences, the MDQFS selects the sentences using compression rate of user's choice.

4 III.

5 Problem Formulation

The automatic text summarization techniques discussed in foregoing section [7][8][9] [12][15] [16]. The major concern of all these techniques is primarily related to text summarization with effective representation of results. But these techniques still have problems as given below:

? They used preprocessed data which diminishes the importance of the proposed method. ? Less number of tags were used while cleaning and summarizing the HTML document. ? Traditional summarization techniques measured the importance of sentence by its location only.

But today, such techniques cannot be adopted in a dynamic environment.

To address these problems an automated frame work for summarizing the search results is proposed in the next section.

IV.

6 Proposed Approach

Proposed framework for Automatic Multiple Document text Summarization using Wordnet and Agility tool is shown in Figure **??**, that takes into account both user query and selection of URLs for summarizing the selected document(s). The whole process, from giving the user query, to getting the summarized results are organized in the following modules.

7 a) Search Engine Interface (SEI)

This module is the heart of the whole system through which user can interact with the proposed system. When user gives a query on the interface of the SE, then SE provides a list of URLs to the end user. The returned results of the SE are stored temporarily.

8 b) Selected Documents (SD)

User can select any number of URLs to be downloaded. These documents are used while sumarizing the document. The SD contains selected and downloaded documents which are selected by the user by using SEI.

9 c) Web Documents Filtration and Code Optimization (WD-FCO)

Web document has been filtered by removing the unwanted HTML tags. These tags are meta, align and CSS style tags etc. Moreover, '&nbsp' has been replaced by space characters as these characters do not contribute to summery generation.

10 d) Topic Block (TB)

A DOM tree is generated corresponding to the filtered document. The leaf nods of this tree are considered as micro block. The micro block of the same parent tag forms a topic block. Therefore, leaf nodes contain the contents of the web page. The topic blocks having the similar information are merged to form a concept block. The concept based similarities are measured by considering the given query keywords, feature keywords, frequency, location of the sentence, tag in which the text appears in the document, uppercase words etc. Step 2. Select the URLs for downloading the WP.

11 User

Step 4. Clean the downloaded web pages.

Step 5. Apply the concept based algorithm [7] for each selected document(s).

Step 6. Select the top scoring sentences for summarization.

Step 6. Returned summarized document to the end user.

Step 7. Stop.

Step 3. Collect the downloaded WP in the local repository.

12 f) Summary Generation (SG)

V.

13 Experimental Setup

The proposed algorithm is implemented in ASP.NET. Apart from this HTML Agility pack for the creation of HTML DOM tree is also used. NUGET software is used for the installation of HTML Agility pack. Moreover, WordNet is used for expressing a distinct VI.

14 Experimental Results

The TB are created from the cleaned document and CB are created from TB. The generated CB is compared and common concept block is chosen for selecting the Featured Keywords (FK). These FK are used to generate the summery of the document(s). The algorithmic view of automatic text summery generation is illustrated by the algorithm given in Figure **??** and description of AMDTSWA frame work is given in Figure 3. concept of a web page. It compares each topic block with other topic blocks and assigned a similarity score. The formation of CB depends upon a thresh hold value. In this article, the topic blocks having the similarity score above 0.5 are merged to form a concept block.

The implemented framework was tested on various web pages of different web sites, but here authors discussed only two of them. These two web sites were www.msit.in and www.piet.edu. Both of these web sites are related to engineering colleges located in New Delhi and Panipat respectively. These web sites were tested on the featured keyword called placement. The obtained summarized results are shown in Figure 4. The summarized results showed the parallel comparison of both selected web sites. This summarized results showed the parallel comparison of both selected web sites. The achieved results contained textual data for normal description. Moreover, summarized results also contained tabular data coming from selected websites. This tabular data contained the information from designated web sites and put it into its own table. From this multi-document summarized result, based on featured keyword any one can easily compare these colleges and can reach to meaningful conclusion.

15 VII.

16 Conclusion

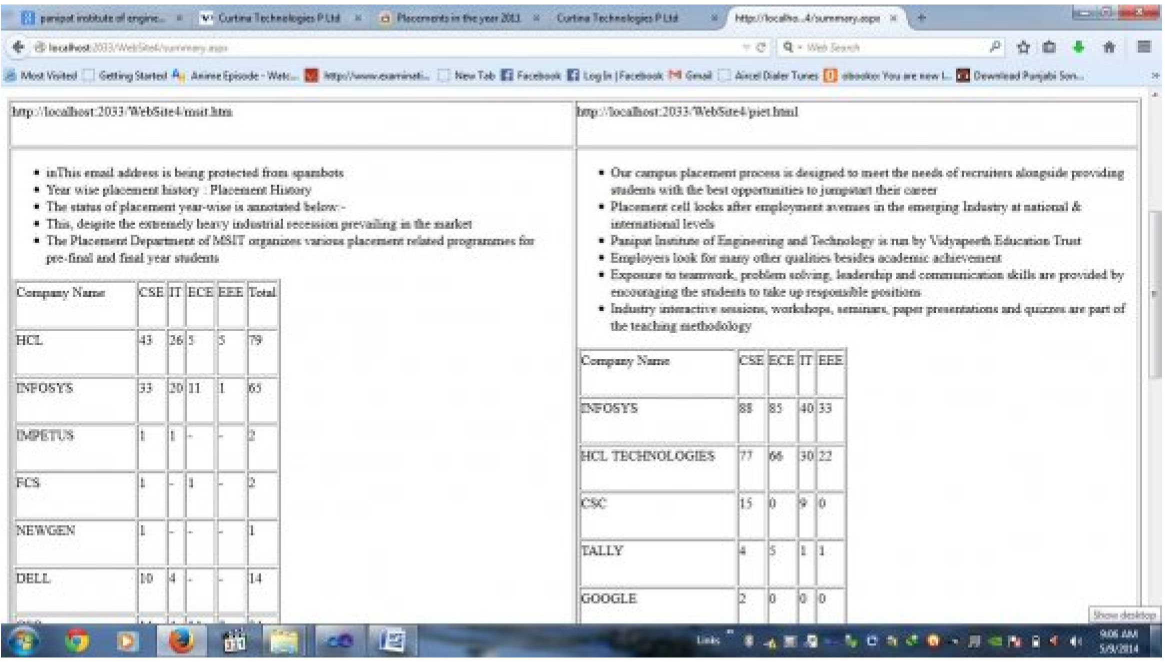
This paper has proposed an automatic text summarization system which can summarize both single as well as multiple documents. The proposed sumarizer system has been implemented in ASP.NET and has been tested. The achieved results have shown that the proposed framework is better than the existing text summarizers in terms of relevancy and presentation of results. The generation of DOM tree and the creation of concept block are done at run time only which removes the need of a static database and saves a lot of memory space needed for storing the contents. Conclusively, by this proposed system of text summarization, the searching and analyzing time of the user is reduced significantly. The comparison of different text summarizers are provided in table1.

17 Summary of Placement for MSIT College

18 Summary of Placement for PIET College

Figure 1:

[1]



http://localhost:2033/WebSite4/msit.htm

- inThis email address is being protected from spambots
- Year wise placement history : Placement History
- The status of placement year-wise is annotated below:-
- This, despite the extremely heavy industrial recession prevailing in the market
- The Placement Department of MSIT organizes various placement related programmes for pre-final and final year students

Company Name	CSE	IT	ECE	EEE	Total
HCL	43	26	5	5	79
INFOSYS	33	20	11	1	65
IMPETUS	1	1	-	-	2
FCS	1	-	1	-	2
NEWGEN	1	-	-	-	1
DELL	10	4	-	-	14

http://localhost:2033/WebSite4/piet.html

- Our campus placement process is designed to meet the needs of recruiters alongside providing students with the best opportunities to jumpstart their career
- Placement cell looks after employment avenues in the emerging Industry at national & international levels
- Panipat Institute of Engineering and Technology is run by Vidyapeeth Education Trust
- Employers look for many other qualities besides academic achievement
- Exposure to teamwork, problem solving, leadership and communication skills are provided by encouraging the students to take up responsible positions
- Industry interactive sessions, workshops, seminars, paper presentations and quizzes are part of the teaching methodology

Company Name	CSE	ECE	IT	EEE
INFOSYS	88	85	40	33
HCL TECHNOLOGIES	77	66	30	22
CSC	15	0	9	0
TALLY	4	5	1	1
GOOGLE	2	0	0	0

12

Figure 2: Figure 1 :Figure 2 :

1

Parameters	Free Summarizer [18]	Auto summarizer [19]	Tools4noobs [20]	MEAD [21]	Comparative [7]	Proposed AMDTSWA
Method used for summary Generation	Extractive	Extractive	Extractive	Extractive	Extractive	Extractive

Figure 3: Table 1 :

[Sornil ()] ‘An Automatic Text Summarization Approach using Content-Based and Graph-Based Characteristics’. Ohm Sornil . *Cybernetics and Intelligent Systems*, 2006. p. . (Print)

[Hovy ()] ‘Automated Text Summarization in SUMMARIST’. Eduard Hovy . 10.3115/ 1119089.1119121. *proceeding of TIPSTER ’98 Proceedings of a workshop on*, (eeding of TIPSTER ’98 Proceedings of a workshop onBaltimore, Maryland) 1999. p. .

[Jung ()] ‘Automatic Text Summarization Using Two-Step Sentence Extraction’. Wooncheol Jung . *Science and Advanced Technology* 2221- 8386. 2005. 2011. springer. 3411 (9) p. .

[Kiani ()] ‘Automatic Text Summarization Using: Hybrid Fuzzy GA-GP’. Arman Kiani , -B . *IEEE International Conference on Fuzzy Systems Sheraton Vancouver Wall Centre Hotel*, (Vancouver, BC, Canada) 2006. p. .

[Svore] ‘Enhancing Single-document Summarization by Combining RankNet and Thirdparty Sources’. Krysta M Svore . DOI: 2007. http://research.microsoft.com/pubs/77563/emnlp_svore07.pdf *Proceedings of the Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, (the Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning) p. .

[Mohamed ()] ‘Improving Query-Based Summarization Using Document Graphs’. Ahmed A Mohamed . *IEEE International Symposium on Signal Processing and Information Technology*, 2006. p. .

[Md and Haque ()] ‘Literature Review of Automatic Multiple Documents Text Summarization’. Md , Haque . *International Journal of Innovation and Applied Studies* 2028-9324. 2013. 3 (1) p. .

[Radev] ‘MEAD -a platform for multidocument multilingual text summarization’. Dragomir Radev . *Proceedings of the 4th International Conference on Language Resources and Evaluation*, (the 4th International Conference on Language Resources and EvaluationLisbon)

[Poonam and Bari (2013)] ‘Multi-Document Text Summarization using Mutual Reinforcement and Relevance Propagation Models Added with Query and Features Profile’. P Poonam , Bari . *International Journal of Advanced Computer Research* (online): 2277-7970. September-2013. (11) p. . (ISSN (print)

[Md and Ali ()] ‘Multi-document Text Summarization: SimWithFirst Based Features and Sentence Co-selection Based Evaluation’. Mohsin Md , Ali . 10.1109/ICFCC.2009.42. *IEEE International Conference on Future Computer and Communication*, 2009. p. .

[ChipraP (2011)] ‘Query Sensitive Comparative Summarization of Search Results using Concept Based Segmentation’. ChipraP . *Computer Science & Engineering: An International Journal (CSEIJ)* 2231 -329X. December 2011. 1 (5) p. .

[Chen ()] ‘Research on Query-based Automatic Summarization of Webpage’. Zhimin Chen . *IEEE ISECS International Colloquium on Computing, Communication, Control, and Management*, 2009. p. .

[Aksoy ()] ‘Semantic Argument Frequency-Based Multi-Document Summarization’. Cem Aksoy . *The 24th International Symposium on Computer and Information Sciences, ISCIS*, (North Cyprus) 2009. IEEE. p. .

[Kumar ()] ‘Summarization of Search Results Based On Concept Segmentation’. Naresh Kumar . *international conference on data acquisition transfer, processing and management (ICDATPM-2014)*, 2014. p. .

[Allan ()] ‘Topic detection and tracking pilot study: final report’. James Allan . *Proceedings of the NAACL-ANLP-AutoSum ’00 Proceedings of the 2000 NAACL-ANLP Workshop on Automatic summ*, (the NAACL-ANLP-AutoSum ’00 the 2000 NAACL-ANLP Workshop on Automatic summ) 1998. 4 p. .