Volume 5, Issue 10, October-2015 ISSN: 2277 128X
International Journal of Advanced Research in Computer Science and Software Engineering
Research Paper
Available online at: www.ijarcsse.com

A Novel Approach for Mining Chosen Keywords Using Text Summarization Extraction System

Sneha Parmar
Computer Engineering Department, SOCET,
Gujarat, India

Abstract*— The objective of text summarization is to decrease the measure of the content while preserving its important information and overall meaning. The ever-increasing user generated digital data available through the Internet has become an important source of information for individuals, organizations and government agencies. And yet, for users to completely find and use that data remains a complex task. Existing popular information access models based on keyword and/or facet searches become less effective in providing access to specific sets of user generated data. In this paper, we present our preliminary development including a new algorithm for keyword extraction and summarization generation simultaneously over a subset of documents.

Keywords*— Text Summarization, Key phrases Extraction, Text mining, Data Mining, Text compression.

I. INTRODUCTION

The purpose of Text Mining is to prepare unstructured data; separate significant numeric lists from the content, and, therefore, makes the data contained in the content accessible to the various data mining algorithms. We mean associations, hypothesis that are not explicitly present in the text source being analysed by the novel information. The websites that contain thousands of web pages in a particular domain or across domains are becoming more and more pervasive due to more information available on the web spaces. These sites allow users to perform keyword searches, and navigate and browse the information through predefined information structures. The large amount of information readily available it is difficult to find correct information in a timely manner these are the common challenges faced by the users. Text summarization is a chain of a compressing a given document into an abbreviated variant by extricating the most imperative information from it.

We say text mining as the revelation by PC of new, previously unknown data, via naturally taking data from a typically huge amount of different unstructured textual resources. Keyword extraction and thought in learning objects is one of the most important subjects in eLearning environments. In this paper a novel model is introduced in order to improve idea in learning objects. The system develops many approaches to solve this problem that gave a high quality result. The model consists of four phases.

The prework phase converts the unstructured text into proper manner.

The first phase of the system removes the stop words, paragraph the text and assigning the POS (tag) for each word in the text and store the result in a tabular form.

The second phase begins from the concentrate the vital key expressions in the content by executing another calculation through algorithm positioning the hopeful words. The framework utilizes the taken key expressions to choose the critical sentence. Every sentence positioned relying upon many features such as the existence of the keywords/key phrase in it, the connection between the sentence and the title by utilizing a typical estimation and other many features.

The Third phase of the proposed system is to taking the sentences with the highest rank.

The Forth phase is the filtering phase. This phase reduced the amount of the candidate sentences in the summary in order to produce qualitative information of previous phases.

A new technique to produce a summary of an original text is investigated in this paper.

II. PRAPOSED SYSTEM IN GENERAL

Text summarization is the procedure of compacting a given archive into a shortened version by extricating the most vital data from it. Approaches for text summarization can be classified into two major categories: extraction and abstraction. The extraction based methodology is to make the summary by extricating the critical sentences from the original report. Though the abstraction based methodology is to develop the summary by rewording concepts of the original document [4]. Procedure of summarization is show in Fig. 1.

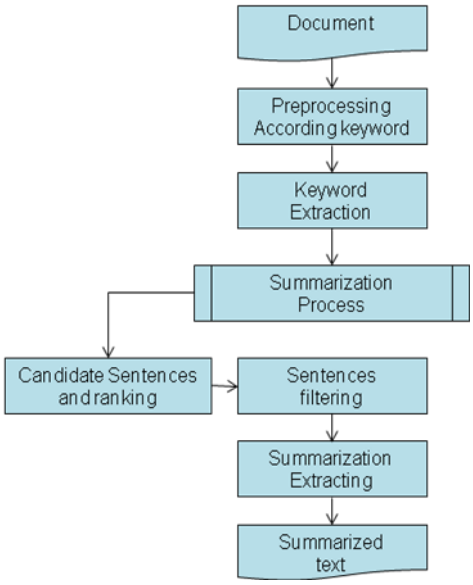

Fig. 1. Process of summarization

Different approaches[3] to automatic summarization works are as follows:

(i) Statistical approach

The summary is created by selecting statistically frequent terms in the document.

(ii) Lexical acknowledgement and classification approach

Selecting sentences based on position in the text report. To start with line in the section and the title are driving possibility to outline the entire report in most of the cases in those text summarization systems.

(iii) Linguistic approach

The proposed framework depends on the word recurrence of the content archive in the wake of dispensing with the stop words which doesn't convey any significance however valuable in sentence.

III. PROPOSED METHOD (ALGORITHM)

In this proposed method mainly we can use Keyword Extraction and Summarization Reinforcement (KASUR)[1] method. The arrangement of keywords extricated from every report is considered as a minimized representation of that record. In light of the co-event of keywords from each record, we can construct a proximity matrix between all sets of record. Utilizing this proximity matrix, we can at that point derive a chart structure W for catching relationship among every one of the clusters and documents. We join this information in a representation application that comprises of a tag cloud of keywords and a system chart appearing connections among clusters and documents. As the user cooperates with the representation interface, the visual representation can assist user with deriving experiences derive insights among the currently returned subset of documents and to aid further exploration of the document collection.

Algorithm:

1. *For each document d from D with set of term T and sentences S.*
2. *Compute $C_{|S|*|T|}$ as a sentence-term frequency matrix where C_{ij} is the frequency that term j occurs in sentence i.*
3. *Compute $R_{|T|*|T|}$ as term-term co-occurrence matrix where R_{ij} is the number of sentences in which term occurs i with term j.*
4. *Compute $E_{|S|*|S|}$ as a sentence-sentence similarity matrix where E_{ij} is the cosine similarity of sentence i and sentence j.*
5. *Compute sentence weight vector $W_{|S|}$ and term weight vector $W_{|T|}$, as following until converging:*

$$W_{|S|} = \theta_1 C_{|S|*|S|} W_{|T|} + \theta_2 E_{|S|*|S|} W_{|S|}$$
$$W_{|T|} = \theta_3 C^T_{|S|*|T|} W_{|s|} + \theta_4 R_{|T|*|T|} W_{|T|}$$

6. *Select top N keywords based on the computed weight.*
7. *Extract salient sentences, remove redundancy and form collection level summary.*

Table. 1. Symbol Description of KASUR

Symbol	Description
d	Document
S	Sentences retrieve from the documents
C	Frequency of sentences
R	Term Co-occurrence matrix
E	Sentence similarity matrix
W	sentence weight vector

IV. PROPOSED NEW METHOD (ALGORITHM)

Proposed New Algorithm:

Input: A Collection of Files

Output: Summery of matching keywords Text

1. *For each document*
2. *Stemming Algorithm:*
   Process to reduce inflected words.
3. *lemmatizing Algorithm:*
   Process of grouping together the different inflected forms of words.
4. *End for each*

Processing Technique:

5. *For each Document*
6. *Extraction of word in to token from given output*
7. *Stop words elimination (using classic method)*
8. *Stemming on extracted output (Using Krovetz or Xerox algorithm)*
9. *Now, Implementing KASUR on the generated output to create the summarization of text;*
10. *Store the location of extracted sentences which are the most suitable for given keywords.*
11. *Check weight of sentences with a given keywords;*
12. *Extract sentences with salient features,*
13. *Remove redundancy of Sentences.*
14. *Form summery with given Scenario*
15. *End for each*
16. *Go to next file and repeat above algorithm;*
17. *End*

V. CONCLUSION

In this Review paper, we concentrated on an automatic text summarization approach by sentence extraction utilizing an enhanced KASUR Algorithm. Ranked sentences are collected by identifying the feature terms and text summary is obtained. This gave the upside of observing the most related sentences to be added to the outline content. The framework delivered the most compacted synopsis with high quality and good results in examination to manual rundown extraction.

REFERENCES

[1] Weijia Xu, Wei Luo, Nicholas Woodward , Yan Zhang, Supporting Data Driven Access through Automatic Keyword Extraction and Summarization IEEE 2015

[2] Rafeeq Al-Hashemi Text Summarization Extraction System (TSES) Using Extracted Keywords International Arab Journal of e-Technology, June 2010

[3] Tulasi Prasad Sariki, Dr. Bharadwaja Kumar, Ramesh Ragala Effective Classroom Presentation Generation Using Text Summarization IJCTA, July-August 2014

[4] M.Suneetha, S. Sameen Fatima Corpus based Automatic Text Summarization System with HMM Tagger International Journal of Soft Computing and Engineering July 2011

[5] Yogesh Kumar Meena, Peeyush Deolia Dinesh Gopalani Optimal Features Set For Extractive Automatic Text Summarization Fifth International Conference on Advanced Computing & Communication Technologies 2015

[6] Eduard Hovy and ChinYew Lin Automated Text Summarization in SUMMARIST

[7] Yoshio Nakao An Algorithm for One-page Summarization of a Long Text Based on Thematic Hierarchy Detection

[8] Penn: Using Word Similarities to better Estimate Sentence Similarity, Sneha Jha and H. Andrew Schwartz and Lyle H. Ungar University of Pennsylvania Philadelphia, PA, USA {jhasneha, hansens, ungar}@seas.upenn.edu

[9] Context-Based Similarity Analysis for Document Summarization S.Prabha, Dr.K.Duraiswamy, B.Priyanga Associate Professor, Department of Information Technology K.S.Rangasamy College of Technology, Tiruchengode – 637215, Tamil Nadu, India, (IJARCET) Volume 3, Issue 4, April 2014

[10] Han Jiawei, Kamber M. *Data Mining: Concepts and Techniques*. San Francisco California: Morgan Kaufmann Publishers, 2001.

