International Journal of Science and Research (IJSR)
ISSN (Online): 2319-7064
Index Copernicus Value (2015): 78.96 | Impact Factor (2015): 6.391

A Survey Paper on Mining Keywords Using Text Summarization Extraction System for Summary Generation over Multiple Documents

Parmar Paresh B.[1], Ketan Patel[2]

[1]G.M.F.E, Himmatnagar G.M.F.E, Himmatnagar

[2]Professor, G.M.F.E, Himmatnagar G.M.F.E, Himmatnagar

Abstract: ***We refer text mining as the discovery by computer of new, previously unknown information, by automatically extracting information from a usually large amount of different unstructured textual resources. Keyword extraction and concept finding in learning objects is one of the most important subjects in eLearning environments. In this paper a novel model is presented in order to improve concept finding in learning objects. The system develops many approaches to solve this problem that gave a high quality result. The model consists of four stages. The preprocess stages convert the unstructured text into structured. In first stage, the system removes the stop words, pars the text and assigning the POS (tag) for each word in the text and store the result in a table. The second stage is to extract the important key phrases in the text by implementing a new algorithm through ranking the candidate words. The system uses the extracted keywords/key phrases to select the important sentence. Each sentence ranked depending on many features such as the existence of the keywords/key phrase in it, the relation between the sentence and the title by using a similarity measurement and other many features. The Third stage of the proposed system is to extract the sentences with the highest rank. The Forth stage is the filtering stage. This stage reduced the amount of the candidate sentences in the summary in order to produce a qualitative summary using KFIDF measurement. A new technique to produce a summary of an original text investigated in this paper.***

Keywords: Text Summarization, Key phrases Extraction, Text mining, Data Mining, Text compression

1. Introduction

Keyword and feature extraction is a fundamental problem in text data mining and also document processing. Majority of document processing applications directly depend on the speed and quality of keyword extraction algorithms. Assignment of high quality keywords manually is expensive and time-consuming. There are various algorithms for automatic keywords extraction that have been recently proposed. Since there is no precise scientific definition of the meaning of a document, different algorithms produce different outputs. Information extraction (IE) is the task of automatically extracting structured information from unstructured and/or semi-structured machine readable documents. A broad goal is to allow computation to be done on the previously processed unstructured data which is used to summarize the text.

In this paper, we talk about Text summarization as a technique which uses keywords as the basic document building block and performs analysis on the document. Text summarization is an important area in Natural Language Processing (NLP).The manual summarization of large documents is a very difficult and time-consuming task; hence there is high demand for fast, effective and reliable automatic text summarization tools and models. This becomes especially important with an exponential growth in the number of electronically available documents on the Internet and enterprise intranets. In presents a sentence reduction system for automatically removing extraneous phrases from sentences that are extracted from a document for summarization purpose. The system uses multiple sources of knowledge to decide which phrases in an extracted sentence can be removed, including syntactic knowledge, context information, and statistics computed from a corpus which consists of examples written by human professionals. Reduction can significantly improve the conciseness of automatic summaries. A new technique to produce a summary of an original text investigated in this paper.

2. The Proposed System Architecture

The following diagram figure.1 represents the proposed system:

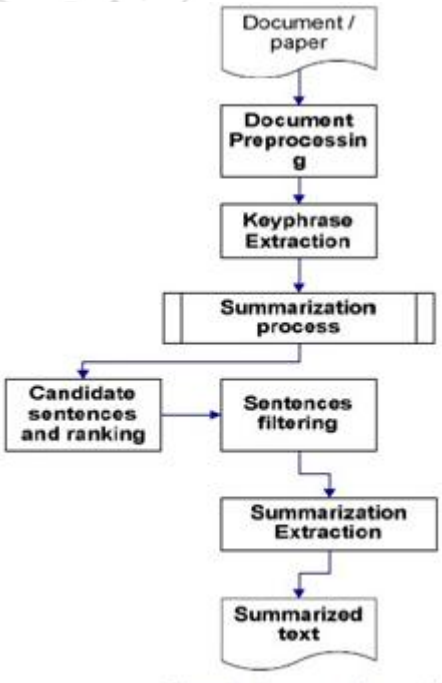

Figure 1. System architecture

The model consists of the following stages:

3. Preprocessing Technique

1. Extraction of words and tokens
2. Stop words elimination
3. Stemming algoritham
4. Implementing Proposed algoritham

Here, Stemming algorithm is used to process repeated words.

Leminotibation
Process to group the different infited words. Together, so that we can analysed as single item. The proposed algorithm will make the process more strengthen and reduce time taken for execution of process. Take multiple document into consideration dataset document for each document from dataset D, with set of team T and Sentence S. There are different approach. All have opted text mining approach. there can be graph mining, multiview leading .

The result will be Top-N keywords extracted from multiple dataset. The pre- processing is a primary step to load the text into the proposed system, and make some processes such as case-folding that transfer the text into the lower case state that improve the accuracy of the system to distinguish similar words. The pre-processing steps are:

3.1 Stop Word Removal

The procedure is to create a filter for those words thatremove them from the text. Using the stop list has the advantage of reducing the size of the candidate keywords.

3.2 Word Tagging

Word tagging is the process of assigning P.O.S) like(noun, verb, and pronoun, Etc.) to each word in a sentence to give word class. The input to a tagging algorithm is a set of words in a natural language and specified tag to each. The first step in any tagging process is to look for the token in a lookup dictionary. The dictionary that created in the proposed system consists of 230,000 words in order to assign words to its right tag. The dictionary had partitioned into tables for each tag type (class) such a stable for (noun, verb, Etc.) based on each P.O.S category. The system searches the tag of the word in the tables and selects the correct tag (if there alternatives) depending on the tags of the previous and next words in the sentence.

3.3 Stemming

Removing suffixes by automatic means is an operation which is especially useful in keyword extraction and information retrieval. The proposed system employs the Porter stemming algorithm with some improvements on its rules for stem. Terms with a common stem will usually have similar meanings, for example:

(Connect, Connected, Connecting, Connection, Connections)

Frequently, the performance of a keyword extraction system will be improved if term groups such as these are conflated into a single term. This may be done by removal of the various suffixes -ED, -ING, -ION, IONS to leave the single term CONNECT. In addition, the suffix stripping process will reduce the number of terms in the system, and hence reduce the size and complexity of the data in the system, which is always advantageous..We will similarity measserement which is a word-distance measurement. Here, Stemming algorithm is used to process repeated words. display the result with the help of repidminor and will implement the proposed algo in Java Programing Language

4. Proposed Algorithm

1) Select desired documents
2) Eliminate stop words
3) Put this result in stemming algorithm
4) Implements decision tree algorithm with feature selection process(Use removal of word sense ambiguity).
5) Find distance between keywords(TF*idF).
6) Keywords with highest similarity will be selected.
7) Add them to the find list.
8) Display result with sentences with selected keywords as summary.

5. Keyphrase Features

The system uses the following features to distinguish relevant word or phrase (Keywords):

- Term frequency
- Inverse Document Frequency
- Existence in the document title and font type.
- Part of speech approach.

5.1 Inverse Document Frequency (IDF)

Terms that occur in only a few documents are after more valuable than ones that occur in many. In other words, it is important to know in how many document of the collection a certain word exists since a word which is common in a document but also common in most documents is less useful when it comes to differentiating that document from other documents.

5.2 Existence in the Document Title and Font Type

Existence in the document title and font type is another feature to gain more score for candidate keywords. Since the proposed system gives more weight to the words that exists in the document title because of its importance andindication of relevance. Capital letters and font type can show the importance of the word so the system takes this into account.

5.3. Part of Speech Approach

After testing the keywords that extracted manually by the authors of articles in field computer science we noted that those keywords fill in one of the following patterns as displayed in table (1). The proposed system improves thisapproach by discover a new set of patterns about (21 rule)that frequently used in computer science. This linguistic

approach extracts the phrases match any of these patterns that used to extract the candidate keywords.

5.4. Keyphrase Weight Calculation

The proposed system computes the weight for each candidate keyphrase using all the features mentioned earlier. The weight represents the strength of the keyphrase, the more weight value the more likely to be a good keyword (keyphrase). We use these results of the extracted keyphrases to be input to the next stage of the text summarization. The range of scores depends on the input text. The system selects N keywords with the highest values.

Table 1. P.O.S. Patterns.

no	POS Patterns
1.	<adj> <noun>
2.	<noun> <noun>
3.	<noun>
4.	<noun> <noun> <noun>
5.	<adj> <adj> <noun>
6.	<adj>
7.	<adj> <adj> <noun> <noun>
8.	<noun> <verb>
9.	<noun> <noun> <noun> <noun>
10.	<noun> <verb> <noun>
11.	<noun> <adj> <noun>
12.	<prep> <adj> <noun>
13.	<adj> <adj> <adj> <noun pl>
14.	<noun> <adj> <noun pl>
15.	<adj> <adj> <adj> <noun>
16.	<noun pl> <noun>
17.	<adj> <propern>
18.	<adj> <noun> <verb>
19.	<adj> <adj>
20.	<adj> <noun> <noun>
21.	<noun> <noun> <verb>

5.5 Classification

The proposed system tries to improve the efficiency of the system by categorizing the document by trying to assign a document to one or multiple predefined categories and to find the similarity to other existing documents in the training set based on their contents.

This process has two benefits one for document classification and the second for feed backing this result to filtering the extracted keywords and to increase the accuracy of the system by discarding the candidate keywords that are irrelevant to the processed document field, since the proposed system is a domain specific.

6. Sentences Selection Features

- Sentence position in the document and in the paragraph.
- Keyphrase existence.
- Existence of indicated words.
- Sentence length.
- Sentence similarity to the Document class.

6.1. Existence of Headings Words

Sentences occurring under certain headings are positively relevant; and topic sentences tend to occur very early or very late in a document and its paragraphs.

6.2. Existence of Indicated Words

By indicated words, we mean that the existence of information that helps to extract important statements. The following is a list of these words:

- Purpose: Information indicating whether the author's principal intent is to offer original research findings, to survey or evaluate the work performed by the others, to present a speculative or theoretical discussion.
- Methods: Information indicating the methods used in conducting the research. Such statement may refer to experimental procedures, mathematical techniques.

6.3. Sentence Length Cut-off feature

Short sentences tend not to be included in summaries. Given a threshold, the feature is true for all sentences longer than the threshold and false otherwise.

7. Post Processing

The system makes filtering on the generated summary toreduce the number of the sentences, and to give more compressed summary.The KFIDF computed for each keyword, the more trivial keyword and frequently used in the class gains more value of KFIDF. Again to the example above if the candidate keyword is neural the filter finds phrases within the system near the word neural then it will select the one which has more KFIDF value

$$\text{KDIDF(w,cat)} = \text{docs(w,cat)} \times LOG\left(\frac{n \times |cats|}{cats\ (word\ )} + 1\right)$$

docs (w,cat)= number of documents in the category cat containing the word w
n- smoothing factor cats(word) = the number of categories in which the word occurs

8. Measurements of Evaluation

Using the generated abstract by the author as the standard against which the algorithm-derived abstract. The results evaluated by Precision and Recall measurements. Precision (P) and Recall (R) are the standard metrics for retrieval effectiveness in information retrieval. They calculated as follows:

$$P = tp / (tp + fp)$$

$$R = tp / (tp + fn)$$

Where tp = sentences in the algorithm-derived list also found in the author list; fp = sentences in the algorithm derived list not found in the author list; fn = sentences in the author list not found in the algorithm derived list. They stand for true positive, false positive and false negative, respectively.

9. Results

Documents from the test set have been selected, and the selected sentences to be in the summary presented in table 2 below:

Table 2. Experiment results.

Text #	Automatic selected	Manual selected	Matched	Precision
1	4	3	3	75%
2	9	8	6	67%
3	8	8	6	75%
4	10	8	6	60%
5	10	10	9	90%
6	14	10	9	64%
7	13	12	10	77%
8	24	22	19	79%
9	30	26	22	73%
10	31	21	13	42%
Overall Precision				70%

10. Conclusion

The work presented her depends on the keyphrases extracted by the system and many other features extracted from the document to get the text summary as a result. This gave the advance of finding the most related sentences to be added to the summary text. The system gave good results in comparison to manual summarization extraction. The system can give the most compressed summary with high quality. The main applications of this work are Web search Engines, text compression and word processor.

References

[1] WeijiaXu,WeiLuo,Nicholas Woodward, Yan Zhang, " Supporting Data Driven Access through Automatic Keyword Extraction and Summarization," IEEE, pp. 1-34, 2015
[2] YogeshKumar,Meena, PeeyushDeolia,DineshGopalani, “Optimal Features Set For Extractive Automatic Text Summarization,” 2015 Fifth International Conference on Advanced Computing & Communication Technologies,2015
[3] TulasiPrasadSariki, Dr.BharadwajaKumar, Ramesh Ragala, “Effective Classroom Presentation Generation Using Text Summarization,” IJCTA , July-August 2014
[4] BrianLott, “Survey of Keyword Ex-traction Techniques,” IEEE-2013
[5] M.Suneetha, S. SameenFatima, “Corpus based Automatic Text Summarization System with HMM Tagger,” International Journal of Soft Computing and Engineering (IJSCE)ISSN: 2231-2307, Volume-1, Issue-3, July 2011
[6] Rafeeq Al-Hashemi , " Text Summarization Extraction System (TSES) Using Extracted Keywords," International Arab Journal of e-Technology, Vol. 1, No. 4, June 2010