International Journal for Research in Applied Science & Engineering Technology (IJRASET)

Automatic Text Document Summarization

Prof. Nihar Ranjan[1], Pranay N. Lonkar[2], Sanket M. Sathe[3], Nayan A. Shendre[4], Sonali M. Shingade[5]
[1,2,3,4,5]*Department of Computer Engineering, SITS, Pune, India*

Abstract— Conventional Information Retrieval approaches are insufficient for the increasingly vast amounts of text data. Typically, only a small amount of the many available documents will be applicable to a given individual or user. Users need tools to compare different documents, rank the importance and relevance of the documents, or find patterns and trends across multiple documents. The amount of information available electronically is growing. So there is increasing demand for automatic method for text summarization. Text summarization is process of reducing the size of text document while preserving its information content. We are going to work on automated creation of summaries of one or more text documents using similarity measure algorithm – cosine similarity.

Keywords—Text Summarization: Tokenization, Stop Words, Stemming, Term Frequency, Inverse Document Frequency

I. INTRODUCTION

A huge amount of data is available over internet. A large amount of data is uploaded over internet every day, which causes the availability of bulk data here. That means we have a large amount of data that will successfully match our search. Also we cannot forget the fact that a large amount of data is also present there which is not suitable for our search. In that case searching out a relevant data that meet our requirements is a tedious and time consuming task. We can face two kinds of problems

Searching a relevant document corresponding to our search.

Absorbing maximum amount of information from that bulk data source.

Here we are working for such a technique that will resolve both of these problems i.e. time and information.

Two fundamental techniques are identified to automatically summarize texts, i.e. abstractive and extractive summarization [1]. Complex summarization techniques are generally based on abstraction. It uses computer generated analysis and amalgamation of the source documents into a completely new documents. Not only is the summarized document shorter but also united, legible and comprehensible. Multidisciplinary approaches in information retrieval linguistics, artificial intelligence and machine learning have been applied to achieve the abstractive summarization.

In extraction task, the automatic system excerpts objects from the whole collection, without modifying the objects themselves. Examples of this include key phrase extraction, where the goal is to select individual words or phrases to "tag" a document, and text document summarization, where the objective is to choose whole sentences (without modifying them) to create a short paragraph summary.

Text summarization techniques can also be classified on the basis of volume of text documents available in the text database. *Single-document* summarization can only distill one document into a shorter version, whereas; *multi-document* summarization can condense a set of documents. *Multi-document* summarization can be seen as an enrichment of single-document summarization and can be used for outlining the information contained in a cluster of documents [2, 3].

Though many of the same techniques used in single-document summarization can also be used in multi-document summarization, there are at least five significant differences:

Anti- redundancy methods are needed since the degree of repetition as previously remarked is considerably higher in a group of topically related articles that in a sole article as each article tends to explain the main point as well as necessary shared background.

The group of articles may contain a temporal dimension, typical in a stream of news reports about an unfolding event, in which case later information may override earlier incomplete reports.

The summary size required by the user will typically be much smaller for collections of topically related documents than for single documents requiring a lower compression factor (i.e. the size of the summary with respect to the size of the document set), thereby requiring a far more careful selection of passages.

The co-reference issue presents a greater challenge when entities and facts occur across documents than in a single-document situation.

The user interface will need to address the users' information seeking goals by allowing rapid elective interaction with the summary such as for the purposes of viewing context of a passage within the summary, view related information to the summary passages including the original document and/or single document summaries, and create new related summaries. The

organization of the paper is as follows. Section II presents some background to the work presented and section III discusses related work. In section IV, we discuss the proposed approach of text summarization followed by conclusion and future work in Section V. Section VI is about acknowledgement and lastly, section VII contains all references.

II. BACKGROUND

Currently, the number of documents retrieved by Web Search Engines is already beyond the capacity of human analysis due to the fact that hundreds of pages of search results are generated for most input queries. Thus document retrieval is not enough and we need a second level of abstraction to reduce this huge amount of data the ability of summarization. Automatic text summarization reduces text contents into most important concepts and ideas under a particular context. This technology may be helpful to identity topics, categorize contents, and summarize documents. However, most previous work on automatic text summarization has emphasized on information abstraction and extraction. Some well-known approaches like Term Frequency/Inverse Document Frequency [4], which summarizes a text based on term frequency weight that is allocated to each term, neural network system tor text document summarization, statistical models, and so on, usually rank sentences and select sentences with higher ranking score as the summary.

Semantic similarity [5] is a concept frequently employed in determining the ranking of a term or sentence. Various semantic similarity techniques are available which can be used for calculating the semantic similarity between text documents. Semantic similarity methods are classified into four main categories:

Edge Counting Methods that measure the similarity between two terms (concepts) as a function of the length of the path linking the terms and on the position of the terms in the taxonomy.

Information Content Methods to measure the difference in information content of two terms as a function of their probability of occurrence in a corpus.

Feature based Methods to measure similarity between two terms as a function of their properties (e.g., their definitions) or based on their relationships to other similar terms in the taxonomy.

Hybrid methods that combine the above three mentioned methods for calculating the semantic similarity.

Since one summary needs to be composed from many documents, there are some issues to be considered. A side problem is that of novelty-detection - given an ordered set of documents, summarize the first document and then summarize only the previously unseen information in all subsequent documents. A higher compression is needed. A 10% summary may be sufficient for one document but with, say, ten documents, concatenating the individual 10% summaries will give a text as long as an average document in the collection - too long. A 1% summary is more like what is needed. A straightforward extractive summary is out of the question.

Also, information may be repeated in different documents and a decision will be needed on which of the intersecting sentences should be included. Rhetorical relations between sentences need to be established - does one sentence contradict another or does it say everything the other says and more? Many more relations exist and these can be used to judge which sentence carry the most new information.

III. RELATED WORK

There are many approaches to summarize documents by finding topics of the document first and scoring the individual sentences with respect to the topics. Sentence clustering has been successfully applied in document summarization to discover the topics present in a document collection. However, existing clustering-based summarization approaches are occasionally targeted for both diversity and coverage of summaries, which are believed to be the two key issues to determine the quality of summaries. The focus of the work [6] is to explore a systematic approach that allows diversity and coverage to be undertaken within an integrated clustering-based summarization framework. Cai et al. [7] developed two co-clustering frameworks, particularly integrated clustering and interactive clustering, to cluster sentences and words simultaneously. Co-clustering frameworks are suggested to grant words to play an explicit role in sentence clustering as an independent text object and to grant simultaneous sentence and word clustering. A fuzzy medoid-based clustering approach for query-oriented multi-document summarization, shown in [8], is successfully employed to generate subsets of sentences where each of them coincides to a subtopic of the related topic. For detecting relevant information and averting unnecessary information in the summaries Lloret and Palomar [9] presented a text summarization tool, called compendium. It incorporates the statistical and cognitive-based techniques for detecting relevant information and for avoiding redundant information it uses textual entailment.

A text summarization evaluation technique named AutoSummENG (AUTOmatic SUMMarization Evaluation using N-gram Graphs) is suggested by Giannakopoulos et al [10]; various methods for evaluation are also discussed for the suggested technique. A language- and domain-independent statistical-based method for single document extractive summarization is suggested by Ledeneva et al [11], to produce a text summary by extracting some sentences from the given text. The main problem for generating an extractive automatic text summary is to detect the most relevant information in the source document.

An extractive text summarization algorithm is proposed by Amulfo et al [12], which use n-grams and maximal frequent word sequences as features in a vector space model. A machine learning ranking algorithm is proposed by Amini et al [13] for single document summarization. The use of machine learning techniques for summarization permits one to adapt summaries to the user needs and to the corpus characteristics. A set of features is first used to make a vector of scores for each sentence in a given document and a classifier is trained in order to produce a global combination of these scores. The ranking algorithm also combines the scores of different features but its criterion contributes to reduce the relative disordering of sentences within a document. A Two-step Sentence Extraction summarization system is created and introduced by Jung et al [14].

IV. PROPOSED SYSTEM

The proposed method can be described in seven steps as shown in Figure I.

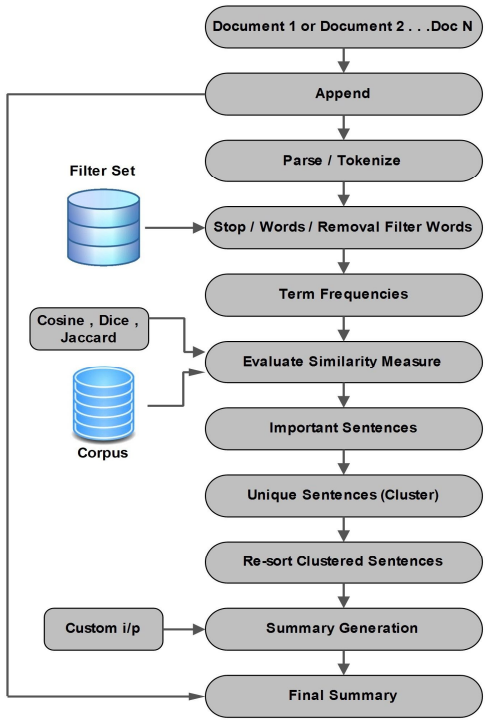

Figure 1: The Proposed System

A. Selection of Text Documents

In the first step text documents which are required to be summarized are given by the user.

B. Append and Tokenization

Text documents are appended and then the file content is tokenized into individual word.

C. Stemming and Removal of Stop Words

We find out the root/stem of a word. Various suffixes are removed; number of words is reduced by having exactly matching stems. Language specific functional words which carry no information are removed.

D. Generation of List of Frequent Words

After eliminating stop words the term-frequent data and inverse document frequency is calculated from text documents and frequent terms are selected which are used to generate text document summary.

E. Sentence Generation

Similarity measure is evaluated using cosine algorithm and important sentences are generated. Unique sentences are clustered and re-sort. Finally, the summary is generated.

F. Update Details in Database

When the summary is generated then its details is stored in the database and is available to the user for information analysis.

G. Setup Web Service

A web service to provide summary of given text documents, will be set up. The Web Service client will send request message consisting of document then the server sends the summary as the response message.

Algo DocSummarizer

Input:

1. Multiple text documents for which summary is to be generated.
2. Value of *N* for generation of *N* lines of summary.

Output:

1. Summary for text documents.
2. Compression Ratio.
3. Retention Ratio.

Steps:

1. Data Pre-processing Phase
   - Retrieve text documents
   - Apply stemming
   - Eliminate stop words
2. For the entire text content
   - Get the TF and IDF
   - Evaluate similarity measure
   - Re-sort the important sentences
   - Add sentences to summary-sentence-list
3. Calculate Compression Ratio and Retention Ratio

Figure 2: Multiple Text Document Summarizer Algorithm

V. CONCLUSION AND FUTURE WORK

Since there is vast amount of textual information available on Web which can't be analyzed by humans, a service oriented approach can be useful for retrieving important information from text documents. In this paper, we developed a text summarizer that is capable of producing a relevantly abstract summary from multiple text documents of same domain as per the number of lines requested by the user. The proposed method is basically an extraction based approach. Our system builds upon previous work in single-document summarization - taking into account some of the major issues arising in multi-document summarization: (i) the need to carefully eliminate redundant information from multiple documents, and achieve high compression ratios, (ii) information about document and passage similarities, and weighting different passages accordingly, and (iii) the importance of temporal information. Future work includes (i) integration of multi-document summarization with document clustering to provide summaries for clusters produced by topic detection and tracking, (ii) generation of coherent temporally based event summaries and, (iii) construction of interactive interfaces so that users can effectively use multi-document summarization to browse and explore large document sets.

VI. ACKNOWLEDGEMENT

We would like to thank our project guide Prof. Nihar Ranjan and all the staff members of our department for giving us valuable time and contributing their experience towards this project.

REFERENCES

[1] Nenkova, Ani, and Kathleen Mckcown. Automatic summarization Now Publishers Inc, 2011.
[2] Canhasi E., Kononenko I. Weighted archetypal analysis of the multi-element graph for query-focused multi-document summarization // Expert Systems with Applications, 2014, vol.41, no.2, pp.535–543.
[3] Luo W., Zhuang F., He Q., Shi Z. Exploiting relevance, coverage, and novelty for query-focused multi-document summarization // Knowledge-Based Systems, 2013, vol.46, pp.33–42.
[4] Salton, Gerard, and Christopher Buckley. "Term-weighting approaches in automatic text retrieval." Information processing & management (1988)
[5] Resnik. Philip. "Semantic similarity in a taxonomy: An information-based measure and its application to problems of ambiguity in natural language." arXiv preprint arXiv: 1105.5444 (2011).
[6] Cai X., Li W., Zhan R. Enhancing diversity and coverage of document summaries through subspace clustering and clustering-based optimization // Information Sciences, 2014, vol.279, pp.764–775.
[7] Cai X., Li W., Zhang R. Combining co-clustering with noise detection for theme-based summarization // ACM Transactions on Speech and Language Processing, 2013, vol.10, no.4, Article 16, 27 pages.
[8] Mei J.-P., Chen L. SumCR: A new subtopic-based extractive approach for text summarization // Knowledge and Information Systems, 2012, vol.31, no.3, pp.527–545.

[9] Lloret E., Palomar M. COMPENDIUM: a text summarization tool for generating summaries of multiple purposes, domains, and genres // Natural Language Engineering, 2013, vol.19, no.2, pp.147–186.
[10] George Giannakopoulos, Vangelis Karkaletsis, George Vouros, "Summarization Evaluation Under an N-Gram Graph Perspective. In View of Combined Evaluation Measures.", TAC2008, (2008).
[11] Yulia Ledeneva, Alexander Gelbukh, and René Arnulfo García-Hernández, "Terms Derived from Frequent Sequences for Extractive Text Summarization", CICLing 2008, LNCS 4919, pp. 593–604, (2008).
[12] Rene Arnulfo García-Hernández, Yulia Ledeneva, "Word Sequence Models for Single Text Summarization", 2009 Second International Conferences on Advances in Computer-Human Interactions, (2009).
[13] Massih R. Amini, Nicolas Usunier, and Patrick Gallinari, "Automatic Text Summarization Based onWord-Clusters and Ranking Algorithms", ECIR 2005, LNCS 3408, pp.142–156, (2005).
[14] Wooncheol Jung, Youngjoong Ko, and Jungyun Seo, "Automatic Text Summarization Using Two-Step Sentence Extraction", AIRS 2004, LNCS 3411, pp. 71 – 81, (2005).