IOSR Journal of Computer Engineering (IOSRJCE)
ISSN : 2278-0661 Volume 1, Issue 6 (Aug-July 2012), PP 01-03
www.iosrjournals.org

A Statistical Approach to perform Web Based Summarization

Kirti Bhatia, Dr. Rajendar Chhillar
[1, 2] *(Department of Computer Science & Applications, M.D University, India)*

Abstract: *Over the past decade more and more users of the Internet rely on the search engines to help them find the information they need. However the information they find depends to a large extent, on the ranking mechanism of the search engines they use. Not surprisingly it in general consists of a large amount of information that is completely irrelevant. Text summarization is a process of reducing the size of a text while preserving its information content. Text Summarization is an emerging technique for understanding the main purpose of any kind of documents. To visualize a large text document within a short duration and small area like PDA screen, summarization provides a greater flexibility and convenience. This research focuses on developing a statistical automatic text summarization approach, K-mixture probabilistic model, to enhancing the quality of summaries. Sentences are ranked and extracted based on their semantic relationships significance values. The objective of this research is thus to propose a statistical approach to text summarization.*
Keywords - *Extraction, Keywords, Statistical approach, Text Summarization, Webpage.*

I. INTRODUCTION

Finding out the information that users need from a large amount of data is a major problem of information retrieval .Search engine is certainly a useful tool for helping users of the Internet find the information they need quickly. Unfortunately, it, in general, consists of a great amount of information that is totally irrelevant. One of the problems is that useful information tends to spread over a large number of similar documents instead of being located in a single document, but it is extremely difficult to identify and retrieve them.

Building a web document summarization system involves Researches in dependence analysis of webs document Clustering, automatic generating summarization and user interface. Most search engines use ranked lists to rank the importance of the return web pages in response to a user query so that the returned information is more relevant to whatever a user is looking for. However, the ranked lists are not summarized in term of topics and are not suitable for browsing task for a very simple reason. The returned information are not classified or categorized. In other words, the returned web pages are interleaved instead of appearing one after another in terms of its category. Thus, users need to waste a lot of time in filtering out all the irrelevant data—even if search engine providers put a lot of time and effort in developing more useful ranking mechanisms.

II. Types of Summaries

Taxonomically one can distinguish among the following type of summaries: Extractive/ non–Extractive generic/query-based, single-document/multidocument and monolingual/ multilingual/cross lingual.

Most existing summarizers work in an extractive fashion, selecting portions of the input documents (e.g. sentences) that are believed to be more salient. Non-extractive summarization includes dynamic reformulation of the extracted content, involving a deeper understanding of the input text, and is therefore limited to small domains while generic summaries attempt to identify salient information in text without the context of a query. The difference between single and multi-document summarization (SDS and MDS) is quite obvious, however some of the types of problems that occur in MDS are qualitatively different from the ones observed in SDS e.g. addressing redundancy across information sources and dealing with contradictory and complimentary information. No true multilingual summarization systems exist yet however, cross –lingual approaches have been applied successfully.

A number of evaluation techniques for summarization have been developed. They are typically classified into two categories. Intrinsic Measures attempt to quantify the similarity of a summary with one or more model summaries produced by humans. Intrinsic measures include precision, Recall, Sentence Overlap, Kappa, and Relative Utility. All of these metrics assume that summaries have been produced in an extractive fashion. Extrinsic measures include using the summaries for a task, e.g. document retrieval, question answering, or text classification.

Traditionally, summarization has been mostly applied to two genres of text: scientific papers and news stories. These genres are distinguished by a high level of stereotypical structure. In both these domains, simply choosing the first few sentences of a text or texts provides a baseline that few systems can better and none can better by much. Attempts to summarize other texts e.g. fiction or e mail, have been somewhat less successful.

III. Relationship Between The Web Document Summarization And Automatic Text Summarization

Automatic text summarization refers to a summary from one or more texts which are highly concise but loyal to express the original text meanings. Correspondingly, Web document summarization is a summary from one or more Web documents which are highly concise but loyal to express original Web document meanings.

The object of automatic text summarization is plain text; the object of Web document summarization is HTML text which includes not only texts, but also: 1) Hyperlinks; 2) pictures; 3) forms or tables; 4) format symbols; 5) other multimedia data. To simplify this study, this thesis excludes the multimedia data of the Web document summarization.

Obviously, devoid of non-textual elements Web document summarization is the same as automatic text summarization.

Automatic text summarization becomes the subset of Web document summarization. With Hyperlink and format symbols as the main features of Web document summarization, the study of which must pay full attention to them besides its textual documents.

IV. Proposed Work

The proposed work is about the summarization of Web Document. The System is statistical based system in which the keyword, phrase etc is extracted and on the analysis basis the summarization task will be performed..To perform the summarization of the web document we need some valid text documents. The complete research work will be performed in following steps :

4.1 Exact Research Document -The first step of research is to extract the web document. For the web document extract we will prefer some news site. We need to perform the web content mining to extract the document.

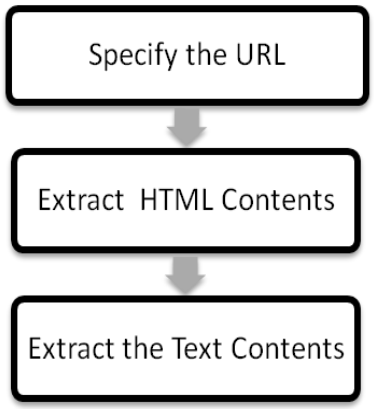

Fig-1

4.2 Document Summary Generation-To summarize a document we need to study and analyze the document in terms of prioritization of keywords, heading, the frequency of the appearance, and the interval of the appearance of word.

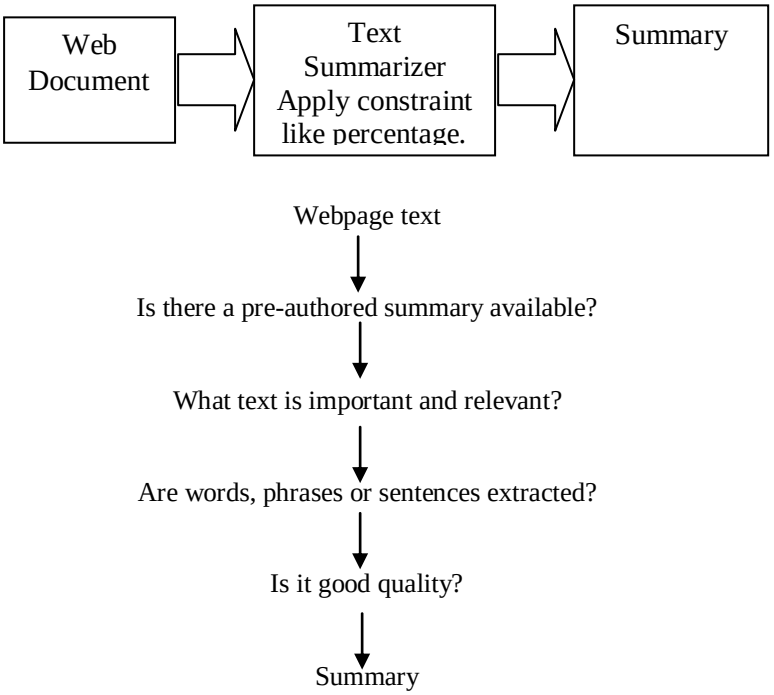

Fig-2

The steps included in the research are given as

- The System will first parses the query language in natural language and finds the major parts in the string.
- Then first it will look for the table and then it parses the string.
- After parsing it will construct the parse tree of the abstracted symbols.
- Once the parse tree is generated will analyze the prioritization and the frequency of the abstracted symbols.
- All these symbols and keywords will be documented in a table.
- Now we will analyze the user requirement of summarization.
- Finally we will extract all the sentences having the same keywords respective to the priority and the user requirement.

4.3 Analysis-Final step of research will be analyzed.

V. Conclusion

In this present work we have defined feature based evaluation approach to perform the document summarization. We have connected the work with web page extraction. In the feature phase, the statistical information is being extracted to perform the summarization.

References

[1] JIANG Xiao-YU," **Improving the Performance of Text Categorization using Automatic Summarization",** International Conference on Computer Modeling and Simulation 978-0-7695-3562-3/09 ©2009 IEEE.

[2] Khushboo S. Thakkar,**" Graph –Based Algorithms for Text Summarization"**, Third International Conference on Emerging Trends in Engineering and Technology 978-0-7695-4246-1/10©2010IEEE.

[3] Munesh Chandra,**" A Statistical approach for Automatic Text Summarization by Extraction"** 2011 International Conference on Communication Systems and Network Technologies 978-0-7695-4437-3/11©2011 IEEE.

[4] LiChengcheng,**"Automatic Text Summarization Based on Rhetorical Structure Theory"**, 2010 International Conference on Computer Application and System Modeling 978-1-4244-7237-6©2010 IEEE.

[5] Jagdish S KALLIMANI,**" Information Retrieval by Text Summarization for an Indian Regional Language",**978-1-4244-6899-7/10@2010IEEE.

[6] Tengfei Ma,"**Multi Document Summarization Using Minimum Distortion",** 2010 IEEE International Conference on Data Mining 1550-4786/10©2010 IEEE.

[7] ZHANG Pei-ying**," Automatic Text Summarization based on sentences clustering and extraction"**, 978-1-4244-4520-2/09©2009 IEEE.

[8] Celal Cigir,**"Generic Text Summarization for Turkish",** 978-1-4244-5023-7/09©2009 IEEE.

[9] Md.MohsinAli**,"Multi-document Text Summarization: SimWithFirst Based Features and Sentence Co-selection Based Evaluation",** 2009 International Conference on Future Computer and Communication 978-0-7695-3591-3/09©2009 IEEE.