International Journal of General
Engineering and Technology (IJGET)
ISSN(P): 2278-9928; ISSN(E): 2278-9936
Vol. 5, Issue 3, Apr – May 2016, 25-28
© IASET

International Academy of Science,
Engineering and Technology
Connecting Researchers; Nurturing Innovations
IASET

GENERATING AUTO TEXT SUMMARIZATION FROM DOCUMENT USING CLUSTERING

JAYA D. KAPOOR[1] & KAILAS K. DEVADKAR[2]

[1]Department of Computer Engineering, Alamuri Ratnamala Institute of Engineering and Technology, Shahpur, India

[2]Department of Information Technology, Sardar Patel Institute of Technology, Andheri, India

ABSTRACT

Auto text summarization is a method of reducing the size of the text document with a software program in a way to generate a summary that retains the most important points of the original document. Interest in the automatic summarization has increased due to the occurrence of information overload, and tremendous growth in quantity of data. Coherent summary can be made using technologies such as considering account variables such as length, writing style and syntax. Google is the one of the good example of the use of summarization technology.

The two technologies viz. Extraction and Abstraction are used for auto text summarizations. Extraction methods work by selecting a content of existing sentences, phrases, or words from the original textual document to form the summary. Unlike, abstractive methods generate an internal semantic of content and then use natural language generation techniques to create a summary that can be related to what a human may generate. Such a summary which may contain words not explicitly present in the original text.

KEYWORDS: Extraction and Abstraction, Coherent Summary, Nearest Neighbour (NN)

INTRODUCTION

Document clustering is very much familiar term in data mining concept and information retrieval system. The said concept was initially used to improve the precision and recall in information retrieval system and also it was used for finding Nearest Neighbour (NN) of the document. This mechanism is also used for organizing the result which is returned by search engines and creating hierarchical clusters within the document. In most of the text processing activities, sentence clustering plays vital role for clustering a complete document. This domain of document clustering mainly focuses on the natural language, artificial intelligence, information retrieval and information extraction [1].

The concept of document summarization is mainly related with the psychology for understanding text and its representations. But the main demand of text summarization technique is simply because of tremendous usage of internet, which has affected the vast use of digital libraries, and data warehouses. And text is a kind of unstructured data that carries different meaning to various different users. Hence we offer a method of auto generation of text.

PROBLEM DEFINATION

Due to the High usage of internet, data warehouses, information organizations, digital libraries, data is growing widely because these are sources which generates textual information that is merely not possible for anyone to handle manually. In much simpler way it can also be said as text is unstructured and has indefinite categories which may carry various different meaning to various different users. Let us say about the task of submitting papers to scientific conference within some proposed categories and tracks which is no more trivial. Most of the authors and writer of scientific articles represents their work using same semantics by using different words or using same words in different method or style. This arises the issues of determining that which paper belong to what category.

EXISTING SYSTEM

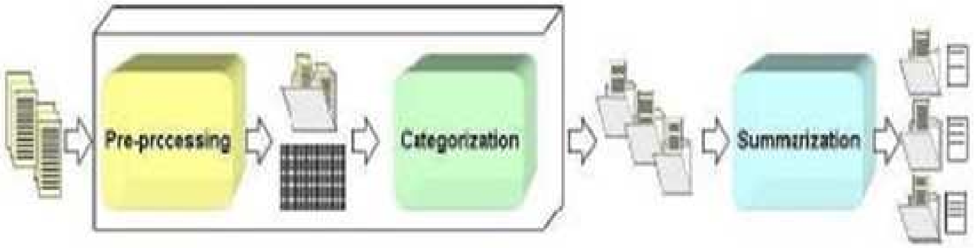

Figure 1

Pre-Processing

Pre-processing is the intial stage of mining activities. This stage allows to collect the document and extract terms systematically from it

Categorization

Categorizations technique plays very important role in handling and organizing the data within the document

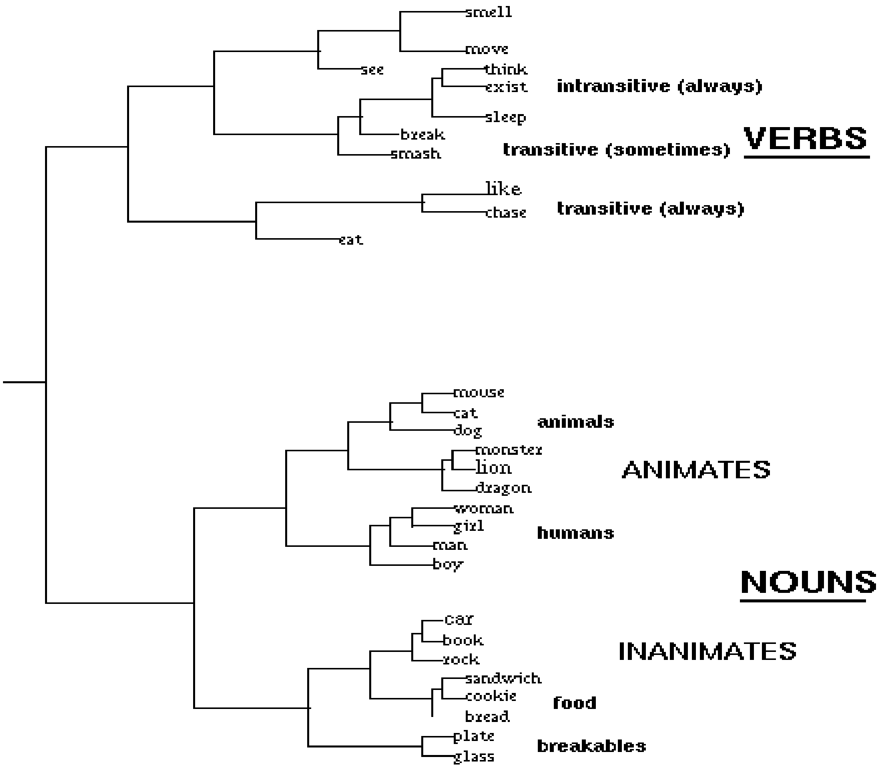

Figure 2

Summarization

The term Summarization is used for retrieval of desired information.

PROPOSED SYSTEM

Extraction-Based Summarization

Extraction Based summarization is used for extracting key-phrases from the document. And this method will help in summarizing document within short paragraph with help of extracting phrases from it where key-phrases are subset of overall content.

Abstraction-Based Summarization

The extraction based summarization simply copy the desired information from the document to create the summary example key phrases, sentences or paragraphs, while abstraction based summarization techniques creates paraphrasing of given document. With respect to extraction technique, abstraction techniques holds the text more strongly. This concept requires the use of natural language technology

Maximum Entropy-Based Summarization

The main goal is to have automatic abstractive summarization within the concept of summarization research but practically most of the systems are based on extractive summarization, because within extractive system the extracted system forms a valid summary within document. Extraction from multi-document summarization mostly depends on the hybrid systems so in this scenario the maximum entropy based summarization comes into the picture. The maximum entropy based summarizations provides high robustness to the system and also increases the quality of the task to be performed.

METHODOLOGY FOR IMPLEMENTATION

To analysis as well to determine which sentences of the document may suits best for auto abstract summarization, we need some calculations by which the information content can be compared with all the sentences within a document. After comparison, a value can be assigned to each sentence according to its quality which can define its significance factor. The significance factor of sentence is achieved from analysis of its words.

Key Phrase Extraction

The key phrase extraction deals with extraction of keywords from the document, that mainly focused on the primary topic of the document/article. Lets take example of research document/article where author manually describe the set of keywords but unlike it if we consider the news paper article, where it would be comparatively more difficult to analysis the keywords. Lets consider the example of news paper article: “the group of army people, rushing to meet president of india’s promise to safe guade the civilian rights in capital” here from this example, an extractive key phrase extractor may select “the group of army people”, “president of india’s”, “safeguard the civilian” as key phrases. These key phrases are directly fetched from the article. Unlike this, abstractive key phrases generates the key phrases by paraphrasing the content within document/article and this generation would be more descriptive in nature from the above example abstractive method would describe it as “political issues” its works similar to a human behaviour of producing the keywords. This method is used in various application and helps in easy browsing through internet by short summaries.

This also helps in improving the informational retrieval by hitting on summary of complete text.

CONCLUSIONS

Hence we proposed that the significant frequency of word can be calculated using its occurrence within the document. We further proposed that significance to the relative position within the words of sentence can give values significant frequency which is useful for calculation for analysis of significant frequency of sentence. Therefore we can state that significant frequency of a sentence depends upon above criteria.

REFERENCES

1. Andrew Skabar, Khaled Abdalgader "Clustering Sentence-Level Text Using a Novel Fuzzy Relational Clustering Algorithm" IEEE Transactions on Knowlegde and Data Engineering vol. 25 No. 1, January 2013
2. Qinru Qin, Qing Wu, Richard Linderman, "Unified Perception- Prediction Model for Context Aware Text Recognition on a Heterogeneous Many-core Platform" International Joint Conference on Neural Networks, san Jose, California July 31- Aug 5, 2011
3. Mohammed Salem Binwahlan, Naomie Salim, "Swarm Based Text Summarization" 2009 International Association of Computer Science and Information Technology- IEEE 2009
4. Shady Shehata, Fakhri Karray, Mohamed Kamel, " An Efficient Concept Based Mining Model for Enhancing Text Clustering" IEEE Transaction on Knowledge and Data Mining vol.22 No. 10 Oct 2010
5. Dingding Wang, Shenghuo Zhu, Chris Ding, "Multi-Document Summarization via Sentence-Level Semantic Analysis and Symmetric Matrix Factorization" SIGIR'08 July 20-24, 2008 – ACM.
6. Martina Naughton, Nicola Strokes, joe Carthy, "Investigating Statistical Techniques for Sentence- Level Event Classification" 22nd International Conference on Computational Linguistics 2008.
7. Furu Wei, Wenjie Li, Qin Lu, Yanxiang He, "Query- Sensitive Mutual Reinforcement Chain and its Appliaction in Query- Oriented Multi- Document Summarization" SIGIR'08 –ACM
8. Wen Pu, Ning Liu, Shuicheng Yan, Jun Yan, Kunqing Xie, Zheng Chen, "Local Word Bag Model for Text Categorization" 7th IEEE International Conference on Data Mining 2007
9. Shady Shehata, Fakhri Karray, Mohamed Kamel, "Enhancing Text Clustering using Concept- Based Mining Model" 6th International Conference on Data Mining 2006 IEEE.
10. Yi Guo, George Stylios, "An Intelligent Algorithm for Automatic Document Summarization" IEEE 2003
11. Antonina K Loptchenko, Jarmo Toivonen, Hannu Vanharanta, "Toward Content Based Retrieval from Scientific Text Corpora" IEEE International Conference on Artificial Intelligence System 2002
12. R. Vasanth Kumar, B. Sankarasubramanium, Dr. S. Rajalakshmi, "An Algorithm for Fuzzy-based Sentence- Level Document Clustering for Micro-level Contradiction Analysis" ICACCI'12- ACM.