iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 2 **Issue:** XI **Month of publication:** November 2014

DOI:

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

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

Auto Text Summarization

Tarun Dugar[1], Rijul Handa[2], Mohit Garg[3], Akshay Gupta[4], Rachna Jain[5]
Student, B.Tech, Department of Computer Science,
Bharati Vidyapeeth's College of Engineering,
New Delhi, India

Abstract— Summaries are an important tool for familiarizing oneself with a subject area. Text summaries are essential when forming an opinion on if reading a document in whole is necessary for our further knowledge acquiring or not. In other words, summaries save time in our daily work. To write a summary of a text is a non-trivial process where one, on one hand has to extract the most central information from the original text, and on the other has to consider the reader of the text and her previous knowledge and possible special interests. Today there are numerous documents, papers, reports and articles available in digital form, but most of them lack summaries. The information in them is often too abundant for it to be possible to manually search, sift and choose which knowledge one should acquire. This information must instead be automatically filtered and extracted in order to avoid drowning in it. Automatic Text Summarization is a technique where a computer summarizes a text. A text is given to the computer and the computer returns a shorter less redundant extract of the original text. So far automatic text summarization has not yet reached the quality possible with manual summarization, where a human interprets the text and writes a completely new shorter text with new lexical and syntactic choices. However, automatic text summarization is untiring, consistent and always available.

Generally speaking there are at least two properties of the summary that must be measured when evaluating summaries and summarization systems - the Compression Ratio, i.e. how much shorter the summary is than the original, and the Retention Ratio, i.e. how much of the central information is retained. This can for example be accomplished by comparison with existing summaries for the given text. Digitally stored information is available in abundance and in a myriad of forms to an extent as to making it near impossible to manually search, sift and choose which information one should incorporate. This information must instead be filtered and extracted in order to avoiding drowning in it.

Keywords— Data Mining, Summary, Scoring, Algorithm

I. INTRODUCTION

With the coming of the information revolution, electronic documents are becoming a principle media of business and academic information. Thousands and thousands of electronic documents are produced and made available on the internet each day. In order to fully utilizing these on-line documents effectively, it is crucial to be able to extract the giz of these documents. Having a Text Summarization system would thus be immensely useful in serving this need. Automatic Text
Summarization is a technique where a computer summarizes a text. A text is given to the computer and the computer returns a shorter less redundant extract of the original text. The basic technology behind this is data mining.

A. Overview of Data Mining

Data mining, *the extraction of hidden predictive information from large databases*, is a powerful new technology with great potential to help companies focus on the most important information in their data warehouses. Data mining tools predict future trends and behaviours, allowing businesses to make proactive, knowledge-driven decisions. The automated, prospective analyses offered by data mining move beyond the analyses of past events provided by retrospective tools typical of decision support systems. Data mining tools can answer business questions that traditionally were too time consuming to resolve. They scour databases for hidden patterns, finding predictive information that experts may miss because it lies outside their expectations.

B. Foundations Of Data Mining

Data mining techniques are the result of a long process of research and product development. This evolution began when business data was first stored on computers, continued with improvements in data access, and more recently, generated technologies that allow users to navigate through their data in real time. Data mining takes this evolutionary process beyond retrospective data access and navigation to prospective and proactive information delivery. Data mining is ready for application in the business community because it is supported by three technologies that are now sufficiently mature:

1) Massive data collection

2) Powerful multiprocessor computers

3) Data mining algorithms

C. Scope Of Data Mining

Data mining derives its name from the similarities between searching for valuable business information in a large database — for example, finding linked products in gigabytes of store scanner data — and mining a mountain for a vein of valuable ore. Both processes require either sifting through an immense amount of material, or intelligently probing it to find exactly where the value resides. Given databases of sufficient size and quality, data mining technology can generate new business opportunities by providing these capabilities:

Automated prediction of trends and behaviours. Data mining automates the process of finding predictive information in large databases. Questions that traditionally required extensive hands-on analysis can now be answered directly from the data — quickly. A typical example of a predictive problem is targeted marketing. Data mining uses data on past promotional mailings to identify the targets most likely to maximize return on investment in future mailings. Other predictive problems include forecasting bankruptcy and other forms of default, and identifying segments of a population likely to respond similarly to given events.

Automated discovery of previously unknown patterns. Data mining tools sweep through databases and identify previously hidden patterns in one step. An example of pattern discovery is the analysis of retail sales data to identify seemingly unrelated products that are often purchased together. Other pattern discovery problems include detecting fraudulent credit card transactions and identifying anomalous data that could represent data entry keying errors.

Data mining techniques can yield the benefits of automation on existing software and hardware platforms, and can be implemented on new systems as existing platforms are upgraded and new products developed. When data mining tools are implemented on high performance parallel processing systems, they can analyse massive databases in minutes. Faster processing means that users can automatically experiment with more models to understand complex data. High speed makes it practical for users to analyse huge quantities of data. Larger databases, in turn, yield improved predictions.

Databases can be larger in both depth and breadth:

More columns. Analysts must often limit the number of variables they examine when doing hands-on analysis due to time constraints. Yet variables that are discarded because they seem unimportant may carry information about unknown patterns. High performance data mining allows users to explore the full depth of a database, without preselecting a subset of variables.

More rows. Larger samples yield lower estimation errors and variance, and allow users to make inferences about small but important segments of a population.

A recent Gartner Group Advanced Technology Research Note listed data mining and artificial intelligence at the top of the five key technology areas that "will clearly have a major impact across a wide range of industries within the next 3 to 5 years."2 Gartner also listed parallel architectures and data mining as two of the top 10 new technologies in which companies will invest during the next 5 years. According to a recent Gartner HPC Research Note, "With the rapid advance in data capture, transmission and storage, large-systems users will increasingly need to implement new and innovative ways to mine the after-market value of their vast stores of detail data, employing MPP [massively parallel processing] systems to create new sources of business advantage (0.9 probability).

The most commonly used techniques in data mining are:

Artificial neural networks: Non-linear predictive models that learn through training and resemble biological neural networks in structure.

Decision trees: Tree-shaped structures that represent sets of decisions. These decisions generate rules for the classification of a dataset. Specific decision tree methods include Classification and Regression Trees (CART) and Chi Square Automatic Interaction Detection (CHAID) .

Genetic algorithms: Optimization techniques that use processes such as genetic combination, mutation, and natural selection in a design based on the concepts of evolution.

Nearest neighbour method: A technique that classifies each record in a dataset based on a combination of the classes of the k

record(s) most similar to it in a historical dataset (where k ³ 1). Sometimes called the k-nearest neighbour technique.

Rule induction: The extraction of useful if-then rules from data based on statistical significance.

II. EXISTING SYSTEMS

A number of Information Extraction Summarization Systems have been developed in specific fields. Even though most of these systems are not currently used in the internet, the potential is great and implementation of such systems in the internet is relatively simple.

Systems have been designed to analyse and summarize medical patient records by extracting diagnoses, symptoms, physical findings, test results and therapeutic treatments. These systems could be used to help health care providers with quality assurance studies. Central databases accessible from different medical centres could be setting up to facilitate transfer of patients as well as for emergencies when the medical records are required immediately. These systems could again be extended to analyse some other databases, example financial statement databases.

A. Case Based Approach

Here the input document is matched against a corpus of relevant and irrelevant texts. Instead of having an explicit set of domain guidelines from a user, the system simply exploits a "training corpus" of representative texts that a user or domain expert has manually classified as either relevant or irrelevant. These predefined representative texts are matched with the document, using statistical techniques to determine the texts that are relevant to the domain of the document. Basically texts that contains only general information are unlikely to be highly correlated with the domain because similar cases will be found in irrelevant as well as relevant texts in the training corpus.Texts that are too specific are also unlikely because there will be very few matching cases. Thus using this statistical technique, only representative texts that contain ley domain-specific information will be extracted. These matched relevant texts could then be used to generate a summary of the text.

Basically, the Case-based approach can be think of as an extension of the basic information extraction system. The problem with the information extraction system is that it retains virtually all the information that is relevant to the domain without any discrimination between important information and details and general information. By including a statistical test in the Case-based approach, we are able to get an idea of the importance and relevancy of the information being extracted for the summary.

B. Document Abridgement

Here a summary is produced by deleting irrelevant texts from the document, retaining only the key passages and sentences of the document. Basically, a typical system consist of two sections, the *Reader* and the *Extractor*. The Reader basically reads in the input text and converts it into internal representations, taking into account the word occurrences and calculating the word weights. The Extractor then determines the particular sentences to be included in the summary by nalysing the word weightings and sentence weightings and then generating the summary from the internal representations.

An example of an experimental system using these methods is the Automatic News Extraction System (ANES) developed by Lisa Rau. This system aims to produce summaries of news from many different sources, had achieved relatively good results in spite of the fact that it is limited by the constraint that it is publication-independent. If developed, this function would prove to be extremely useful for categorising and locating information on the internet by providing summaries to all wide varieties of documents available on the net.

III. PROPOSED SYSTEM

A. SCORING ALGORITHM:

Our program works on following logic:

Word Score

PRIMARY WORD SCORE

Stop Words: These are some insignificant words that are so commonly used in the English language that no text can be created without them. They therefore provide no real idea about the textual theme, and have therefore, been neglected while scoring sentences.

EXAMPLE: I, a, an, of, am, the, et cetera.

Cue Words: These are words usually used in concluding sentences of a text, making sentences containing them crucial for any

given summary. Cue Words provide closure to a given matter, and have therefore, been given prime importance while scoring sentences.

Example: Thus, hence, summary, conclusion, et cetera.

Basic Dictionary Words: 850 words of the English language have been defined as the most frequently used words that add meaning to a sentence. These words form the backbone of our algorithm, and have been vital in the creation of a sensible summary. We have hence, given these words moderate importance while scoring sentences.

Proper Nouns, Numbers and abbreviations: Proper Nouns in most cases form the central theme of a given text. Albeit, the identification of proper nouns without the use of linguistic methods was difficult, we have been successful in identifying them in most cases. Proper Nouns provide semantics to the summary, and have therefore been given high importance while scoring sentences.

FINAL WORD SCORE

Word Frequency: Once basic scores have been allotted to words, their final score is calculated on the basis of their frequency of occurrence in the document. Words in the text which are repeated more frequently than others contain a more profound impression of the context, and have therefore been given a higher importance.

Sentence Score:

Primary Score: Using the above methods, a final word score is calculated, and the sum of word scores of a sentence gives a sentence score. This gives long sentence a clear advantage over their smaller counterparts, which might not necessarily be of lesser importance.

Final Score: By multiplying the score so obtained by the ratio "average length / current length" the above drawback can be nullified to a large extent, and a final sentence score is obtained.

Optimizations

List of basic words, cue words etc are delivered in the form of a text file along with the software. Contents of these text files are stored in the memory at runtime using 'Maps' (in C++).

The entered text has been stored into two types of Data Structures:

- Linked List
- Maps

Special consideration has been given to words written in quotes because many times these words convey important facts.

Generally first sentence from a text presents main points about the content of the entire text. So words appearing in the initial sentence are considered more important.

A special list of titles such as Mr, Mrs, Sir, Capt etc is also considered while extracting words from a sentence.

A special file called 'log.txt' is created for synchronization between C++ and VB during the runtime.

Important Features

1. GUI is implemented, which allows the user to directly input text or select a file from his system, and get output displayed on the screen itself.

2. Output is written in a new file called '<original file name>_summary.txt', it's default location is same as the input file. User can also decide the destination of the output file.

3. If a user inputs the text by himself a file called 'article.txt' is created in the 'summary' directory. Now this file is treated as the input file.

4. Logic is implemented using C++ and GUI is implemented in Visual basic.

5. An installer file is available for easy distribution of the software.

ALGORITHM DESCRIPTION

1. Open the input file, if it has some content then go to next step otherwise a warning is popped.
2. List of basic words, cue words and title words are loaded into the memory using maps.
3. Extract the first sentence from the given text files and creates a map of the words present in this sentence.
4. Extract every sentence from the file and store it in a linked list.
5. For every extracted sentence extract words from them, filter the extracted words and pass them through various word.
6. Before going for the next sentence compute the total score of this sentence. This is done by using sentence scoring techniques.
7. Sort the sentences in descending order based on their final scores.
8. Select 50% sentences from the above list and sort them in ascending order based on their sentence number.
9. The final list of sentences is the summary of the given text.

IV. CONCLUSION

In this paper, a new technique of auto text summarization is proposed in which uses a scoring algorithm. In this process the score of each word is calculated using the method describes in the algorithm and then certain words are chosen to make the summary.

Advantages of the proposed method:

A. Some more scoring techniques, based upon the placement of sentences could have been implemented.
B. Reading from Word documents and HTML files could have been implemented.
C. Natural Language processing could have been implemented

FUTURE SCOPE

The possibilities in this project are endless. With the development of Natural Language Processing (NLP), the following don't remain mere thoughts...

A. Generating newspaper headlines, given the article.
B. Filling up forms, given text containing the necessary data.
C. Creating a bio-data, from a textual detail if the person.

REFERENCES

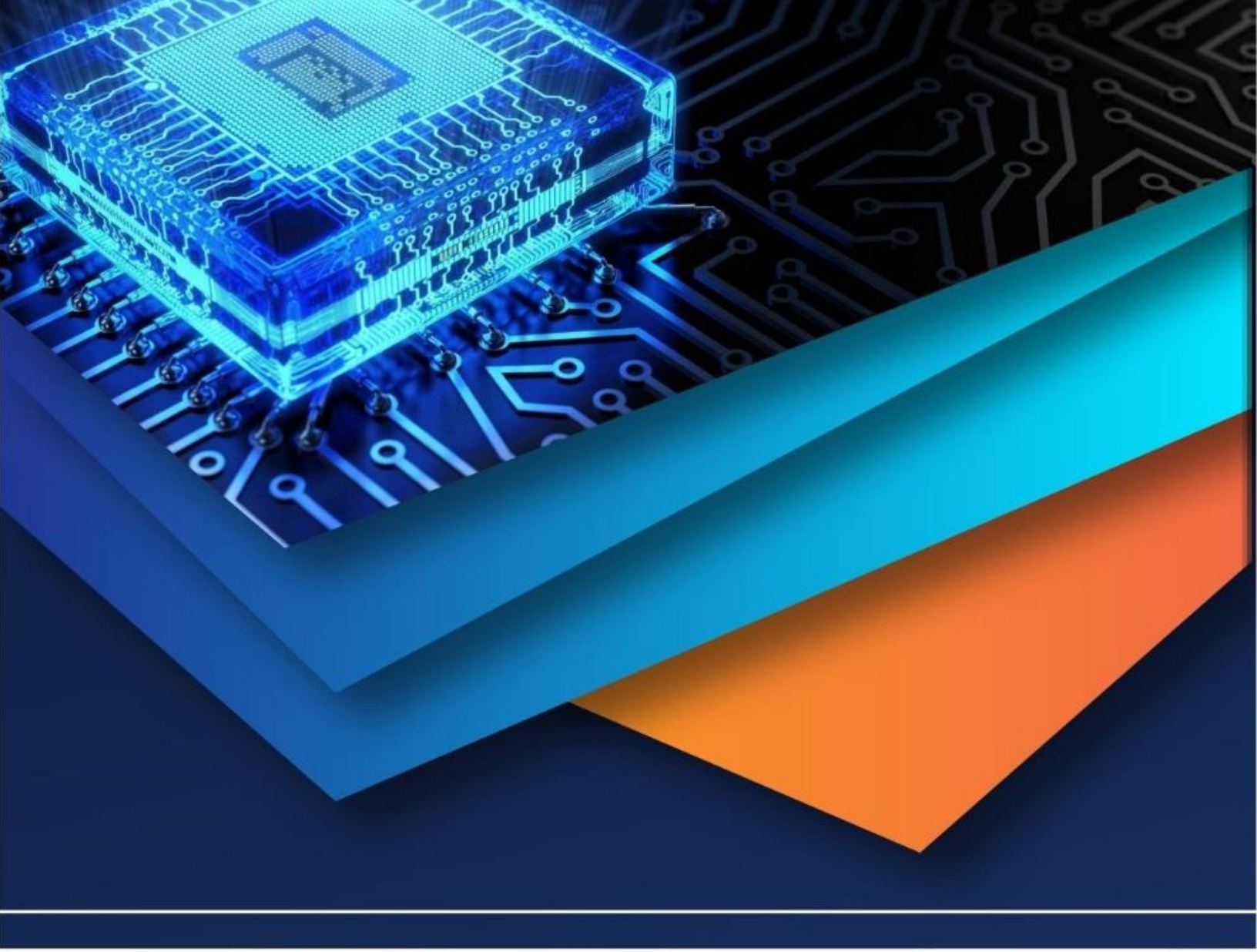
[1]Amy J.C Trappey, Charles V .Trappey, "An R&D Knowledge Management method of patent document summarization", Industrial Management & Data System, vol 108 pp -245-257,2008.

[2] Edmundson, H. P(1969) “New method in automatic Extracting” journal of ACM 1969, 16(2): 264-285.

[3] Erkan G., and Radev, D. R., “LexRank: Graph-based Lexical Centrality as Salience in Text Summarization”, J. Artif. Intell. Res. (JAIR), 22, pp. 457-479, 2004.

[4] Goldstein, J., Kantrowitz, M., Mittal, V., Carbonell, J., 1999. Summarizing text documents: sentence selection and evaluation metrics. In: Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR’99), Berkeley, CA, USA, pp. 121–128. Hahn, U., Mani, I., 2000. The challenges of automatic summarization. IEEE-Computer 33 (11), 29–36.

[5] Websites : Google , Wikipedia.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY