AN IMPROVED AUTOMATIC CONTEXT SUMMARY GENERATOR SYSTEM USING DATA-MINING

Ephraim O. Atonuje and Annie O. Egwali

atonujeephraim@gmail.com, annie.egwali@uniben.edu

Department of Computer Science, Faculty of Physical Sciences University of Benin, Benin City, Nigeria.

ABSTRACT

This study is designed to develop a context summary generator system using data mining technique. Despite the growing availability of electronic documents and the accessibility of desktop publishing technology, abstracts are still manually generated. In the present paper, an automated system that can produce a summary extract of any document fed into it is developed so that a reader can take a look at the contextual interpretation of what the given document states. The system is developed using NET Framework, MS Access, a relational database and the NET high level programming language. It is established that its application would enable readers save time and effort in finding useful information from particular articles or documents. Short versions of lengthy sentences are generated by the system using summarization techniques, while attempting to maintain their meaning.

Key words: Context summary, data mining, automated system, electronic documents, NET Framework, MS Access, NET high programming language.

INTRODUNCTION

In recent decades, the introduction of internet services, information technologies and use of effective data servers have resulted in the generation of large quantity of data and information online. Hesabi *et al*,. (2015) opined that most of the information resulted from social network, sensor sources, cloud storage and others. This usually leads to the problem of how to manage and put to use some aspects of the resulting massive data for human use. When data is too large, storage for use as well as the ability to retrieve them for analytical operations, become real difficult, time consuming and energy sapping.

The setbacks of documents being read and summarized manually cannot be over emphasized. Documents may be processed in large quantity, mostly in educational circles where they are read and reviewed so as to understand the content of the documents. Some factors that are responsible for making the manual reading of documents tedious include; the lot of effort involved in reading a whole document and understanding the important points from it and the fact that, if we have to allow employees to work on hundreds of documents, this may lead to errors.

The solution to all of these problems is to summarize the massive data into a compact form and still retains the original meaning. Automated data summarization (ADS) is an aspect of data mining which focuses on producing or finding a sample superset of recorded observations which give readers knowledge about a huge volume of useful information about an entire population.

ADS is of immense benefit to both individuals and large data processing companies and also suitable for application in industries and organizations everywhere. The authors observed that key information is needed to be read from huge quantity of texts such as novels, newspapers and legal documents. As such, ADS is used for generation of concise text. Some of its techniques include summarization of videos, images and documents. Camargo and Gonzalez (2009) opined that summarization of documents is aimed at creating a piece of the original document by discovering the sentences which have the most relevant information, while in image summarization, the system tries to pick out the most remarkable pictures. For recordings, the system tries to pin point a significant event from uneventful ones.

By automatic summarization, we mean the reduction of content of any document with the use of a software application to produce a summary which usually possess important aspect of written information. Abderrafih (2010) described it as a technique which can produce an understandable summary concentrate on flexible items like structure, sentence and style to text in a word processor and length. Hoplaros *et al,*. (2014) presented four metrics which could be used to characterize data summarization results, proposed two additional metrics, interestingness and intelligibility and used the proposed metrics to evaluate existing summarization techniques on well- known network traffic data sets.

A major change which has occurred in the availability of information to users in recent time is the fact that social networks, cloud collections, online stories and textbooks have provided a large volume of written literature from which researchers, journalists and other users are exposed to tap. This change however, has not provided corresponding techniques which prevent users from generating abstract manually and this makes it time consuming and energy sapping

Although a lot of work has been done on summarization, (see the important papers of Altmani and Menai (2022), Agrawal *et al.*, (2019), Al Saeid *et al*, (2018), Bengio *et al*, (2003), Chem and Zhuge (2014), Cohan and Goharian (2018), Gambhir and Gupa (2017) and Gupta and Gupta (2019)), it appears that articles on the application of data mining to summarization appear to be scanty in existing literature, which is the motivation of this research work.

In the present paper, an automated context summary generator system is designed using data mining technique. The system utilizes NET framework, MS Access, a relational database and the NET high level programming language, which can produce a summary extract of any document fed into it. The generator structure involves the development of a model for extracting detailed content from a document to be used for summarizing a huge volume of information available, design an application that will collect documents as input and apply the model to create an output that is a summarized abstract of the document. To justify the effectiveness of our technique, an application was presented.

MATERIALS AND METHODS

The technique employed here for data summarization is referred to as data mining. Communication enterprises, retailers as well as financial houses employ data mining in

fixing their product prices, positioning of their products and determination of customers' desires for their services.

Data Mining:

This is the act of finding or discovering correlations, insights and patterns in data ware houses/ large volumes of information through categorization techniques, usually at the point where statistical data, huge database and machine learning appear common.

DATA MINING METHODOLOGY

Apriori Algorithm (Rao and Gupta, 2012) will be used for the extraction of the contents of the documents which will be used for the generation of the abstract.

The algorithm will be used to find the frequency of the most commonly documented collections of terms in sentences documentation.

The text in the documents will be divided into units of sentences. Then the sentences are tested for the most commonly occurring word cluster combination. The primary n sentences with the main occurrences of the common word-groups are used because the phrases that make up the abstract, wherever n is the number of phrases that make up the abstract. This can be explained with an application as follows:

Example:

We put into consideration Sentences which appeared in a document as displayed in Table 1 below:

Table 1. Document Item Set

Sentence	**Value**
Sentence 1	Circuit connections appear in parallel and series.
Sentence 2	We will connect in parallel and series.
Sentence 3	Televisions exemplify parallel.
Sentence 4	Solar-panels exemplify recent type series.

We realize the most common word pairs within the sentences using the Apriori algorithm.

Tabular Presentation of the Algorithm

Steps	**Algorithm**
1.	Get words which will be sorted
2.	Set a value s for the most frequency size (In this instance, $s = 2$).
3.	Start pass- one through the items.
4.	After the completion of pass- one, look for the count of every item.
5.	If the count of the item is greater than or equal to s i.e. Count (item i) >= s, then the item i is frequent. Save this for next pass.
6.	When passing a pair of ends, check the count of each item attempt.
7.	If greater than or equal to s, the pair is taken into account to be frequent, i.e. Count (item i, item j) >= s.

The outcomes in Table 2 below were obtained after making the first and with the aid of the preceding algorithm.

Table 2: Item/Word Frequency Table:

Item (word)	**Frequency**
Are	2
Recent	1
And	2
In	2
Complex	1
Connections	1
Parallel	3
Series	3
Televisions	1
Appear	1
Type	1
Circuit	1
Connect	1
We	1
Will	1
Solar-panels	1

Now we get the words that occur the most number of times, higher than s value, that is = 2 and then display them as in Table 3 below:

Table 3: Frequent Word/item frequency table:

Item (word)	**Frequency**
Parallel	3
Series	3
Exemplify	2
And	2
In	2

The frequent words in the table above are written in pairs and presented in Table 4 as follows:

Table 4: Item pairs/Frequent Word:

Item / Word Pairs
Parallel Series
Parallel Exemplify
Parallel And
Parallel In
Series Exemplify
Series And
Series In
Exemplify And
Exemplify In
And In

Next, the frequencies of the most occurring word pairs, which occur in the initial document sentences (see in Table 1) are counted and presented in Table 5 below:

Table 5: Frequency table of the Most Occurring Item (word) Pairs:

Item (word)	**Frequency**
Parallel Series	2
Parallel Exemplify	1
Parallel And	2
In	1
Series Exemplify	1
Series And	2
Series In	1
Exemplify And	0
Exemplify In	0
And In	1

Now we collect the pairs with the highest frequencies, those that are equivalent to the value of s, that is, = 2 or those above 2 and represent them in Table 6 below:

Table 6: Frequency Table of Paired frequency Items/Words:

Item (word)	**Frequency**
Parallel Series	2
Parallel And	2
Series And	2

Finally, sentences with the highest frequency combination of item pairs are presented in Table 7 as displayed below:

Table 7: Frequency Table of the Most Occurring Word Pairs:

Sentence	Value	Paired Words	Paired Word Frequency
Sentence 1	Circuit connections appear in parallel and series.	(Parallel Series) (Parallel And) (Series And)	3
Sentence 2	We will connect in parallel and series.	(Parallel Series) (Parallel And) (Series And)	3
Sentence 3	Televisions exemplify complex parallel.	-	0
Sentence 4	Solar-panels exemplify recent type series.	-	0

It is easily seen from Table 7 that sentences 1 and 2 have the highest frequencies of pair words that is 3 each. Hence the sentences with serial number 1 and 2 will be used as a combination of sentences that will form the abstract. This is so because the two sentences with serial numbers 1 and 2 possess most of the words that can give meaning to the focus of the document. As such, a combined pair of the sentences 1 and 2 will make up the abstract, that is, "This book is about cats and dogs. We will talk about dogs and cats".

PROBLEMS AND WEAKNESS OF MANUAL SYSTEM OF DATA SUMMARIZATION

The setbacks of manually sorting out documents to create a summary of the text are highlighted below.

- Reading and figuring out the key points throughout a text requires a lot of energy and time.
- A huge amount of manpower is needed to read through as well as extract essential excerpts from a report efficiently and can result in high expenditure by the company or entity handling the document processing.
- If a few hands are engaged to handle hundreds of documents, a lot of processing errors may occur and even as at that, the time of completion of the work may be delayed.

THE FUNCTIONALITY OF THE PROPOSED SYSTEM DESCRIPTION

For the Abstract Extractor Software to be developed, the system will make use of a programming language called, Visual Basic NET object-oriented language.

The system will be developed with the following features: (1)Provide authentication details for the security of the documents from unauthorized users. (2) Collect text document as input. (3) Provide facility to store the documents in the database. (4) Enable the number of sentences that the abstract document will be

made up of. (5) Generate the abstract document based on the Apriori Algorithm. (6) Display the generated abstract to user. (7) Give the user the option to save the document. (8) Enable the user to open and view the saved documents and abstracts.

ARCHITECTURE OF THE SYSTEM

The system architecture is modeled in the diagram below.

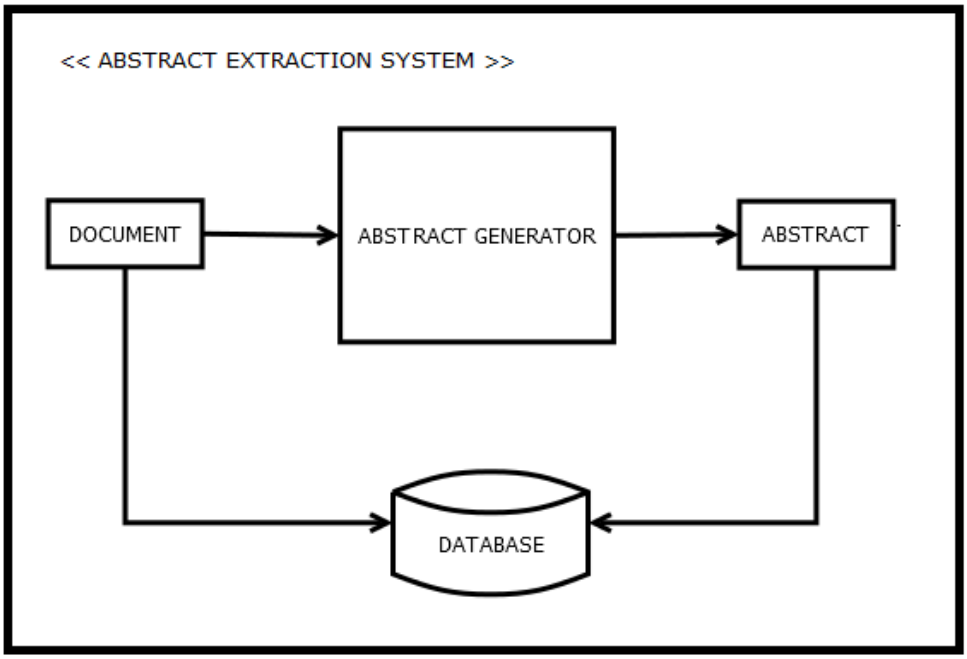


Figure 1 System Architecture

The system will accept inputs in the form of documents. These documents can be stored in the database. The document is passed into the abstract generator where the extraction process occurs using the Apriori Algorithm. The result is the Abstract document which is sent out as output. This can be stored in the database.

PROPOSED SYSTEM ALGORITHM

The system algorithm that will be used for the extraction of the abstract from the document is presented as follows:.

1. Start. 2. Choose the File. 3. Extract the phrases. 4. Set the size of Abstract L. 5. Get a word combination frequency within the sentence. 6. Get L number of phrases with the strongest frequencies of word mix. 7. To generate the abstract text, connect sentences.
8. Display the abstract text. 9. End.

ENTITY RELATIONSHIP DIAGRAM

The entities that make up the proposed system, their attributes and the relationship shared between the attributes are displayed in the figure 2 below.

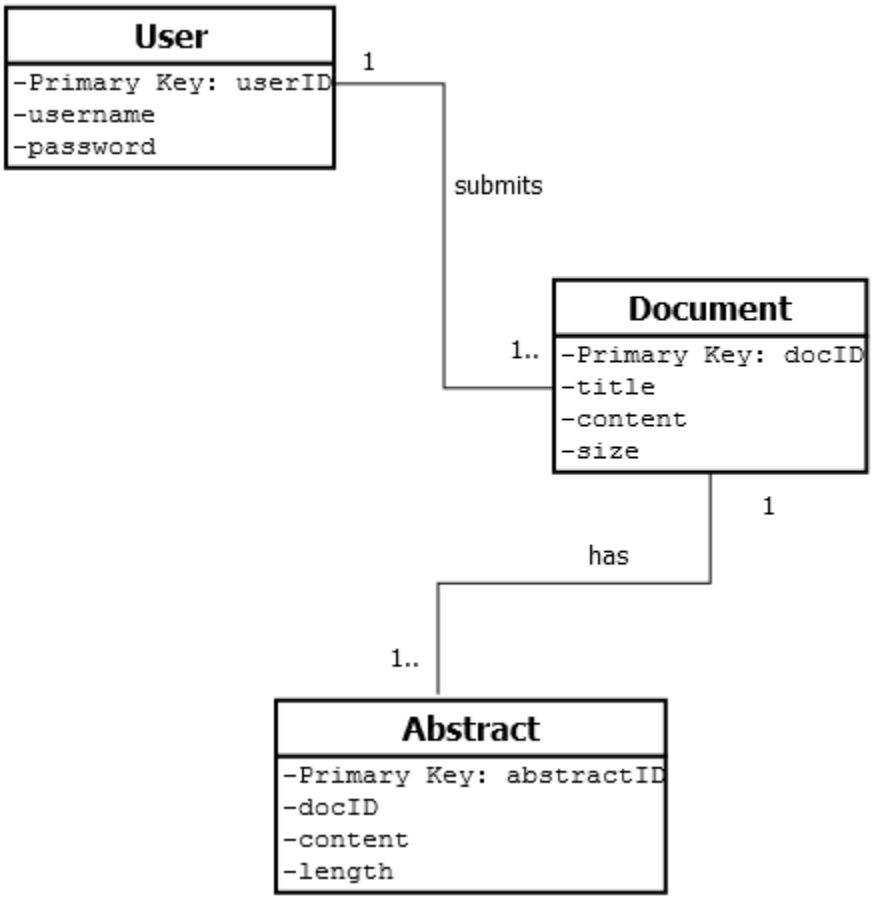


Figure 2 Entity Relationship Diagram

Users Function Performance Diagram

Often called '**the use case diagram**,' it describes the users of a system together with the functions that they can perform on the system. The use case diagram for the present system is displayed in figure 3 below:

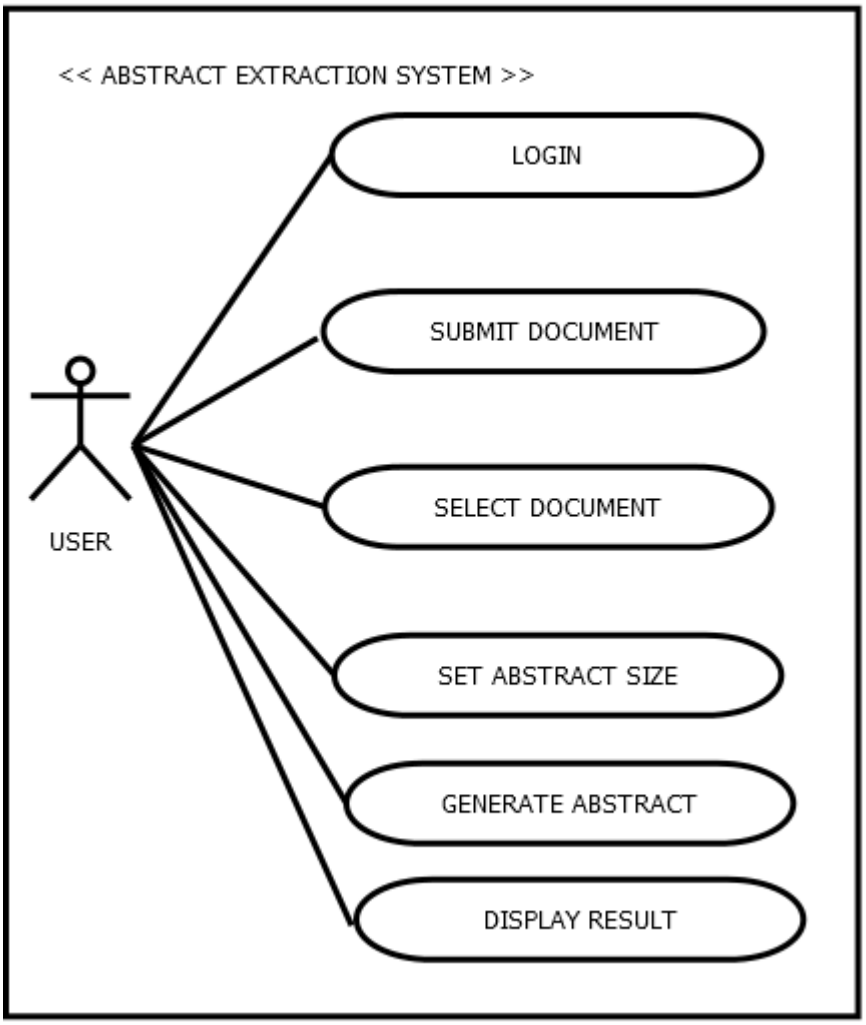


Figure 3: Use Case Diagram

Program Flowchart

The program flowchart showing how the program will be run for document extraction is illustrated in the diagram below.

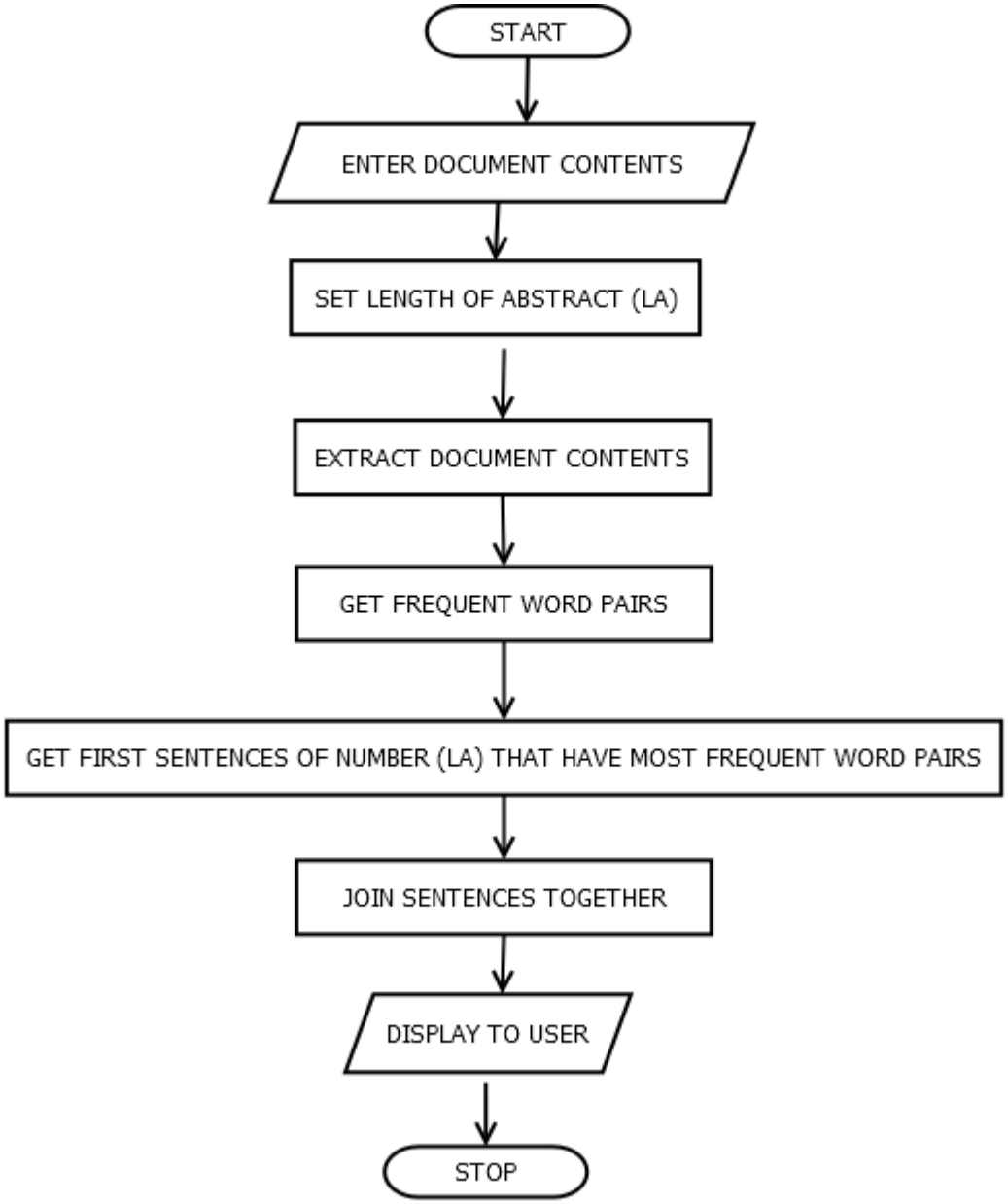


Figure 4: Program Flowchart

SYSTEM FLOWCHART

The system flowchart showing how the users of the system will interact with the interfaces of the system is illustrated in the diagram below.

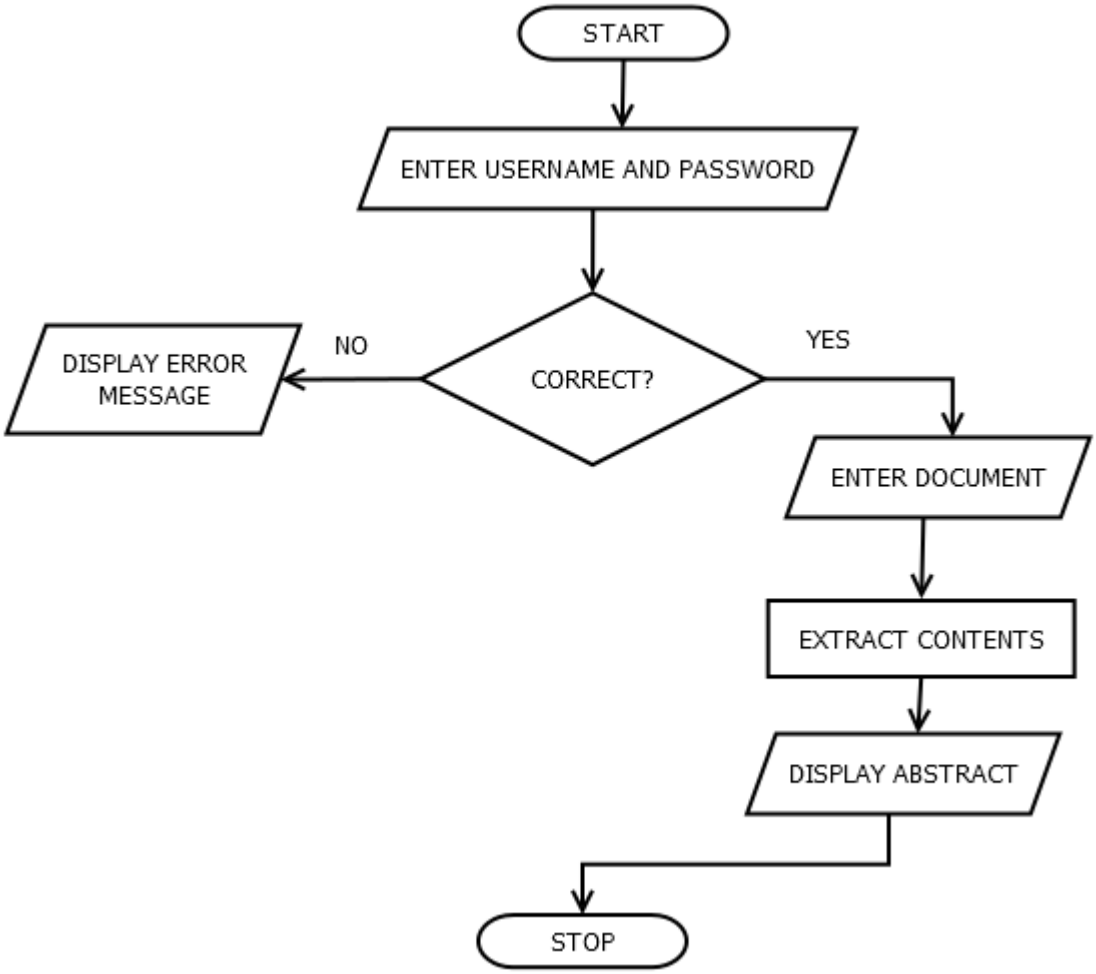
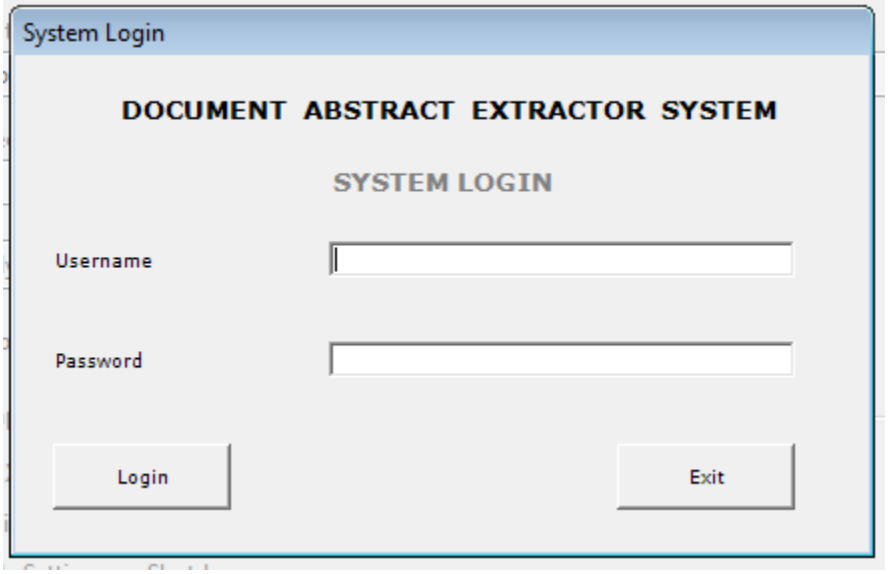


Figure 5: System Flowchart

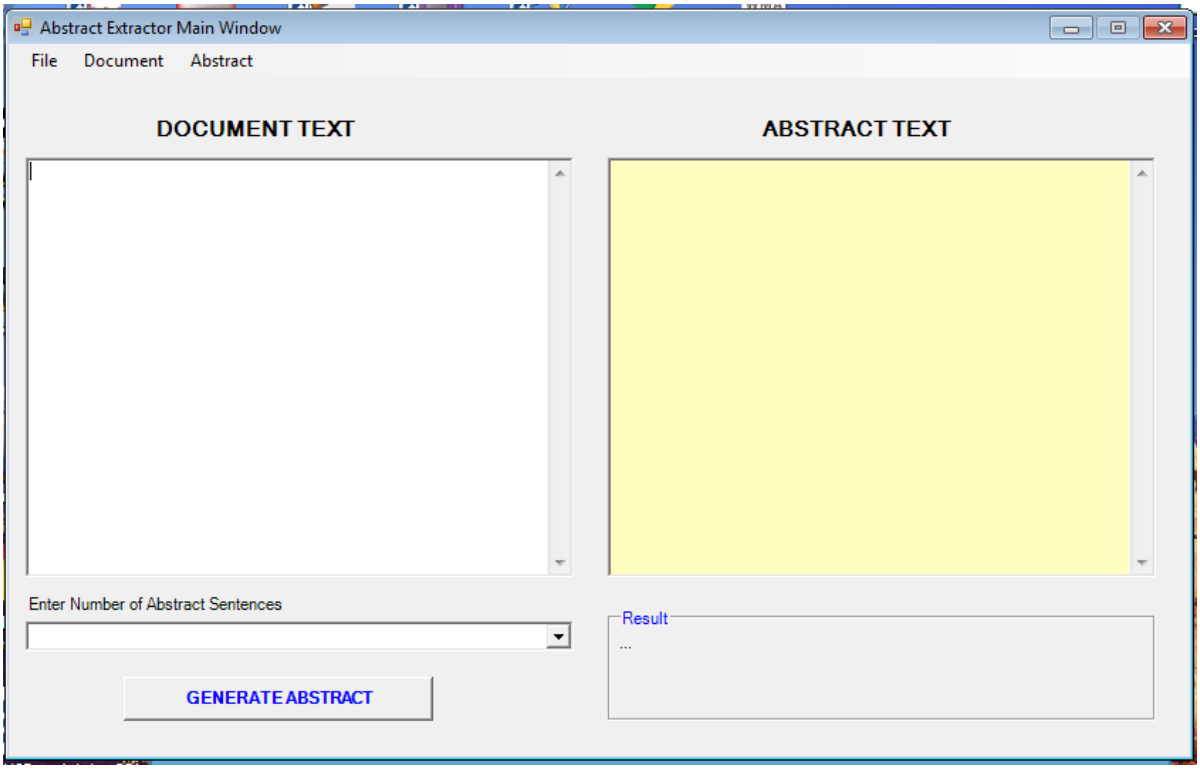
APPLICATION AND DISCUSSIONS

The application of the proposed Document Abstract Extraction Program is implemented using a login module and the main program module which comprises the system's abstract module as presented below:

Login Module: The module below gives a user access to the software.



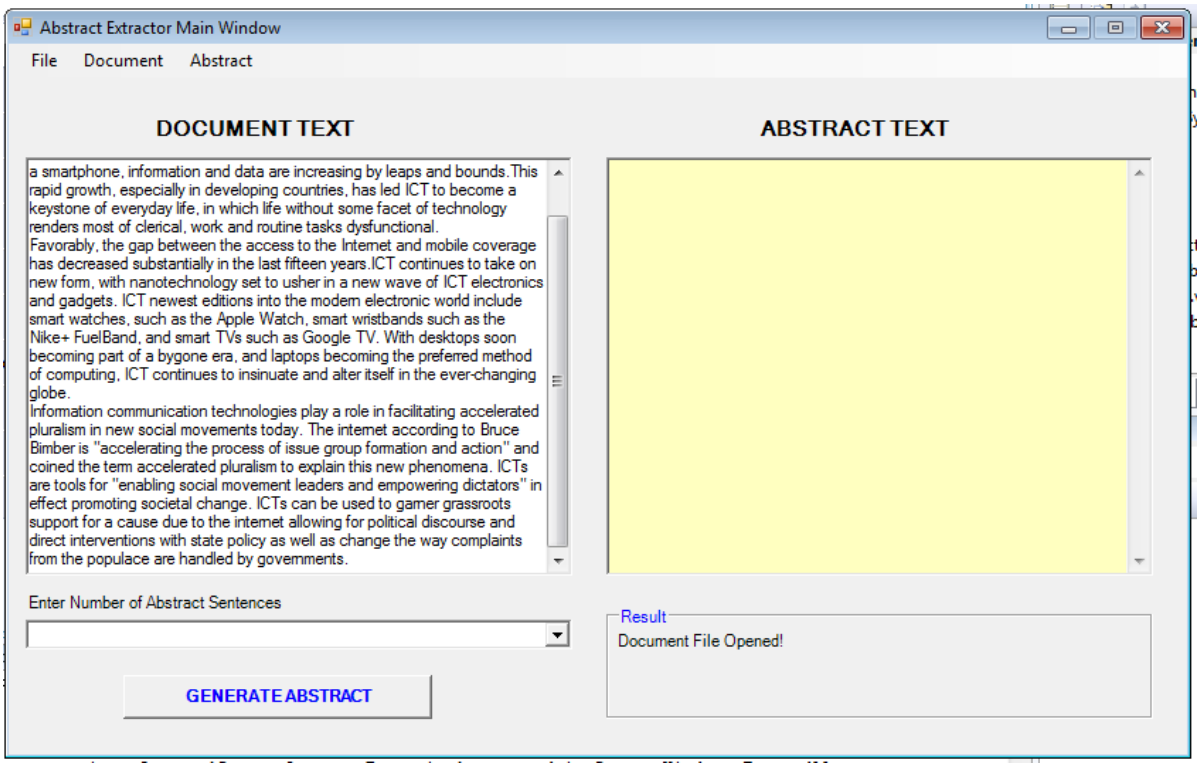
Main Program Module: Below is the system's abstract module.



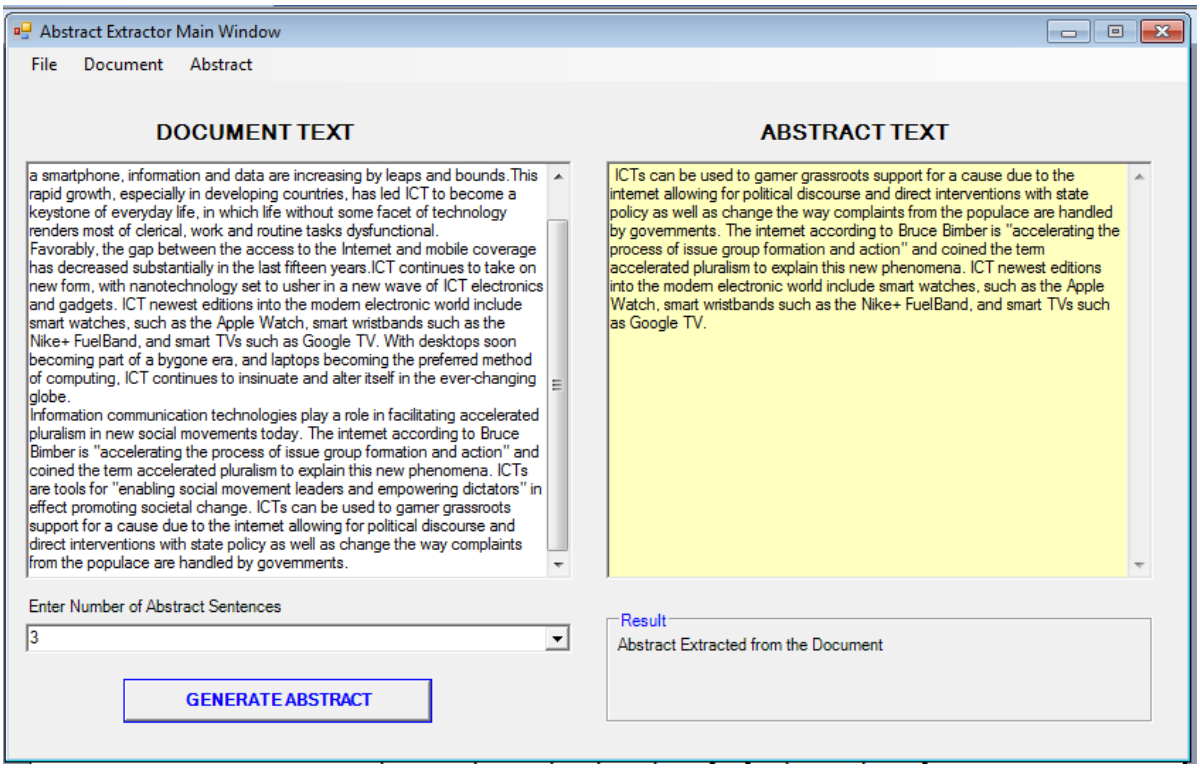
The text volume of text from which context summary is to be extracted usually appears at the document text column whereas the resulted summary appears at the abstract text column.

SYSTEM EVALUATION

In the present section, we display the outcome of assessment of the abstract extraction structure of the document. Document Entry is done by presenting a document entry textbox and indicating how documents can be entered into the box. The format appears as below:



The results of summarized text that are expected depend of the number of abstract sentences chosen by the user of the proposed system. For example, if the user intends to obtain a data summarization containing three abstract sentence document extractions, the outcome of the abstract extraction having been completed is specified to be three (3) with the outcome displayed on a main module screen in an abstract textbox. This type is displayed in the abstract textbox below:



Document Extraction with 3 as the Number of Abstract Sentences

The above tests show that the documents have been successfully extracted and also the number of sentences can be regulated by the researcher.

CONCLUSION

In the system design, the current system was analyzed and the new system described using Unified Modeling Language (UML). It involves the application flowcharts together with system architecture. Individual units of database as well as their properties which were to be employed to grasp data were precisely defined. The building process as well as the proof of reality or testing of the document extraction system was presented along with test case scenarios. The results of the test showed that the system was working as were required.

The purpose of the present article was to provide a software system for data extraction to summarize content in any documentation. This was achieved by using an algorithm of data mining aimed at determining merging of sentences and phrases representing large volumes of documents' essential notions. This system was developed using the association of VBN together with the Access relational database structure.

Creating abstracts from documents required that the documents are read through by the data processor and the main concepts of the documentation taken and then used to describe the nature of the document. The important texts that highlight the essence of the documentation must be highlighted in the summarized version of the abstract. The overall contribution of this article to knowledge is the minimization of the

amount of time and energy involved in summarizing abstracts from large documents and being able to describe the concepts laid out in the documentation without loss in meaning.

REFERENCES

Agrawal, K., Mittal, A. and Pudi, V. (2019). Scalable Semi- Supervised Extraction of Structured Information from Scientific Literature. Proceedings of the Workshop on Extracting Structured Knowledge from Scientific Publications, (2019), 11 – 20.

Altmami, N. I. and Menai, M. E. (2022). Automatic Summarization of Scientific Articles: A Survey. Journal of King Saud University – Computer and Informatioin Sciences, 34(4), 1011 -1028.

Al Saied, H., Pugue, N. and Lamirel, J. (2018). Automatic Summarization of Scientific Publications using a Feature Selection Approach. Int. Jour. Digital Lib. (2018), 1 – 13.

Bengio, Y. Ducharme, P. V. and Jauvin, C. (2003). A Neutral Probabilistic Language Model. Jour. Machine Learn. Res. 3(2003), 1137 – 1155.

Chem, J. and Zhuge, H. (2014). Summarization of Scientific Documents by Detecting Common Facts in Citations. Future Generation Comp. Syst. 32(2014), 246 – 252.

Cohem, A. and Goharian, N. (2018). Scientific Document Summarization via Citation Contextualization and Scientific Discourse. Int. Jour. Digital Lib. 19(2-3), 287 – 303.

Cormode, G. and Muthukrishnan, S. (2005). An improved Data Stream Summary: the Count-min Sketch and its Application. Journal of Algorithms, 55(1), 58 -75.

Dunning, T. (2021). The t-digest: efficient Estimates of Distributions. Software Impact, 7, Article 100049.

Gambhir, M. and Gupta, V. (2017). Recent Automatic Text Summarization Techniques: A Survey. Artificial Intelligence Review, 47(1). Doi 10 1007/s10462 -016 – 9475 -9.

Gupta, S. and Gupta, S. K. (2019). Abstractive Summarization, An Overview of the State of the Art. Expert syst. App. 121 (2019), 49 – 65.

Hesabi, Z. R; Tari, Z; Goscinski, A. and Fahad, A. (2015). Data Summarization Techniques for big Data – A Survey. Doi: 10. 1007/978 – 1- 4939 – 2092 – 1. 38 researchgate.net.

Hoplaros, D., Tari, Z. and Khali, I. (2014). Data Summarization for Network Traffic Monitoring. Journal of Network and Computer Applications, 37, 194 – 205.

Jain, A. K. and Dubes, R. C. (1998). Algorithms for Clustering Data. Prentice – Hall Inc.Engle Wood Cliff, New Jessey.

Jaiwei, H., Micheline, K. and Jian, P.(2011). Data Mining: Concepts and Techniques (3rd Ed.), Morgan Kanfmann ISSN 978-0-12.381479-1.

Kaabneh, K. (2005). Web-based Digital Video Sequencing System Journal of Computer Science, 1(2), 221 – 224. Doi.10.3844/icssp

Rao, S. and Gupta, R. (2012). Implementing Improved Algorithm Over Apriori Data Mining Association Rule Algorithm. International Journal of Computer Science and Technology, 3(1), 489 – 493.

Wang, H. and Wang, S. (2011). Ontology-based Data Summarization Engine: A Designed Methodology. Journal of Computer Information Systems, 53(1), 48 – 56.