Chapter - III

Text Communication Approach

3.1 Introduction

Electronic messages or the mails get saved in the data servers. Mail service providers limit a storage space per user account. They also limit one email attachment size.

The reviews claim, a huge amount of data is getting created daily about 2.5 quintillion bytes [46]. To meet this criterion huge and advanced resources are essential. The storage media is getting completely shifted on the digital storage formats after 2000 [47]. The increase in average data upload on and download from internet has been increasing at a substantial rate [48].

Mail transfer over the internet is increasing at a tremendous speed. It could be understood by the statistical information that the number of emails sent in the year 1995 was recorded as 100 billion; in 2002 this number reached 5.5 trillion; in 2010 this number went across 294 billion [49]. A huge growth is apparent through this data. On an average, 20% time out of total internet access is used in searching through emails and files. Stubbing is a mechanism to share links with the receiver. These are the links to different files images, video streams, etc. This can be a method of reducing the load on email servers.

Another issue is the compulsion of proprietary software use. Documents with formatting by some proprietary software change the file format with its release. The users get locked into a system of buying software though they don’t want change.

A document with the formatting details and computer information can be much bigger to a simple text document. Mail servers provide limited space to their user’s email accounts. The internet communication speed is increasing steadily [50]. The bigger files need more time for transfer. There is a need of a cost-effective application that reduces size of the file. So that it can be easily stored in the mail servers. These files can be accessed and searched at a high speed. A solution to the large file size issue can be of using compression techniques. There are number of compression algorithms categorized in two types; lossy and lossless [51, 52]. The later algorithms are used for compressing text data [53]. As the attachments of the email are lot in number, they need huge memory space. The average memory space required by a user is increasing at a fast rate [54].

In this proposed work an adaptive BIT-STREAM mechanism is proposed. This method extracts from documents having formatting information. The mechanism proposed uses stream readers for extracting characters. The C#.Net development language has the stream readers. The extracted characters are temporarily stored. These characters are attached with the mail servers for communication.

LZ-77, LZ-78 and LZW are some examples of compression algorithms based on pointers and dictionary principle [55, 56].

3.2 Representation of Text in computers

Text in computers is the representation of characters, symbols, digits. Characters combined properly form meaningful words; sequentially as per grammar if arranged, result in paragraphs. Journalism started with textual communication. Text is stored or represented in computers by different formats. American Standard Code of Information Interchange (ASCII) is one of the popular text formats used for storing the information. It has basically a chart indicating a unique code value for each character to be represented. As shown in the figure below, the 1 and 0 are the representation values of the character codes and the voltage values that get stored in the memory.

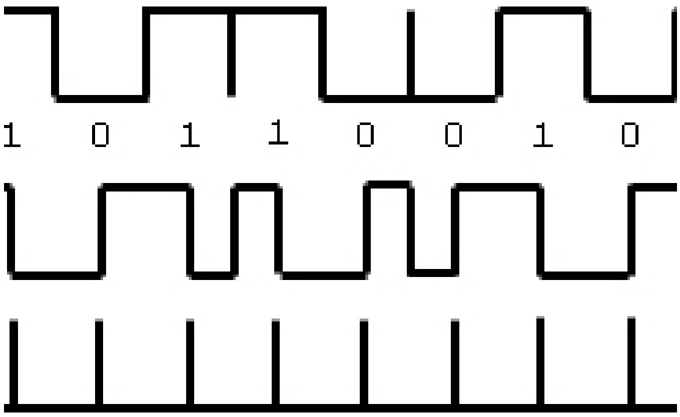


Figure 3.1: Representation of memory values

3.3 Electronic Representation of Text

In the digital electronic representation of characters, formatting information, drawings, digits are stored using numbers. These numbers are unique that are used in storage purpose and display purpose. These numbers are used by computer processing mechanism for identification of the character associated with it. The association of codes has been following standard procedures defined by the consortium. For Accuracy and uniqueness the standard code association with characters is same with all computers. So that, whenever any information is to be shared between distant users, intended characters are represented using these associated codes. Most of the email servers use the same ASCII codes for information exchange. Hence, the text representation of information received through email is same on all computers.

The standard QWERTY keyboards are used by computers for data entry into computer systems. The keyboard has a unique, special arrangement of keys over it. The programs, software or applications that we want to use, associate the specific codes with the keys available on the keyboard. These codes enable data entry. The specific and unique codes have the shape information of each character, the way it is going to be displayed.

Different font styles have different display representation of characters. Font styles are associated with specific codes that represent different character sets. Though all font style codes are coded with identical codes, the displayed shape of the character will be different than other style.

To meet the requirement of display shapes of different character sets of different languages, some advances and special methodology is required for association of codes to the characters. In multilingual scenario, this becomes crucial. One available method for representation to cater large character sets is Unicode. It can be observed with character sets of languages that, eight bit character codes are sufficient for languages having small alphabet set. They can be easily incorporated to the text having individual letters and punctuation. Text formatting software allows use of appropriate fonts for specific codes. The textual input using this software is supported with encoding information about the text inputs.

3.4 Paradigm of Electronic Representation

The method of association of codes to display characters has special concern with

- **Binary language**

The forms of the signals used in memory of computers are of the binary systems. The information represented in the computers gets stored in the primary and secondary memory in the binary language. The language is comprised of two symbols for symbolic representation. The symbols represent the voltage value stored physically in the memory. The text information having characters are associated with the unique character codes. These codes are numeric so; they are converted for the binary storage values.

- **Files, folders**

The continuous flow of characters or information is segmented in the form of files in the memory for ease of retrieval and separation. The files are the units of a different and separate content. The files may be interlinked with other files. Files may have references to other files. With the advances in technology, the embedding of objects of different types in a single file has also become possible. These files can further be kept in different compartments called as folders.

- **Data and Meta Data**

The information stored inside the memory is the content that a user wants to store, retrieve, edit, etc. The storage software or the applications used for storage has the data to be stored and some additional information. This additional information is the Meta data which is the information about the data.

- **Text Formats**

At the base level, the text is categorized as the proprietary and non-proprietary formats.

- **Non-proprietary**

The open and free file formats that don't need any proprietary software for accessing them are non-proprietary text formats. These formats can be opened or accessed with simple text editors. There is no compulsion to use any pre-owned software.

- **Proprietary**

The text formats which can be accessed with only proprietary software are the proprietary text formats. These file formats have a compulsion of using the particular file handling or formatting software.

Text formatting includes features like page settings, font settings, color settings, layout settings, etc.

3.5 Introduction to Bit-Streaming

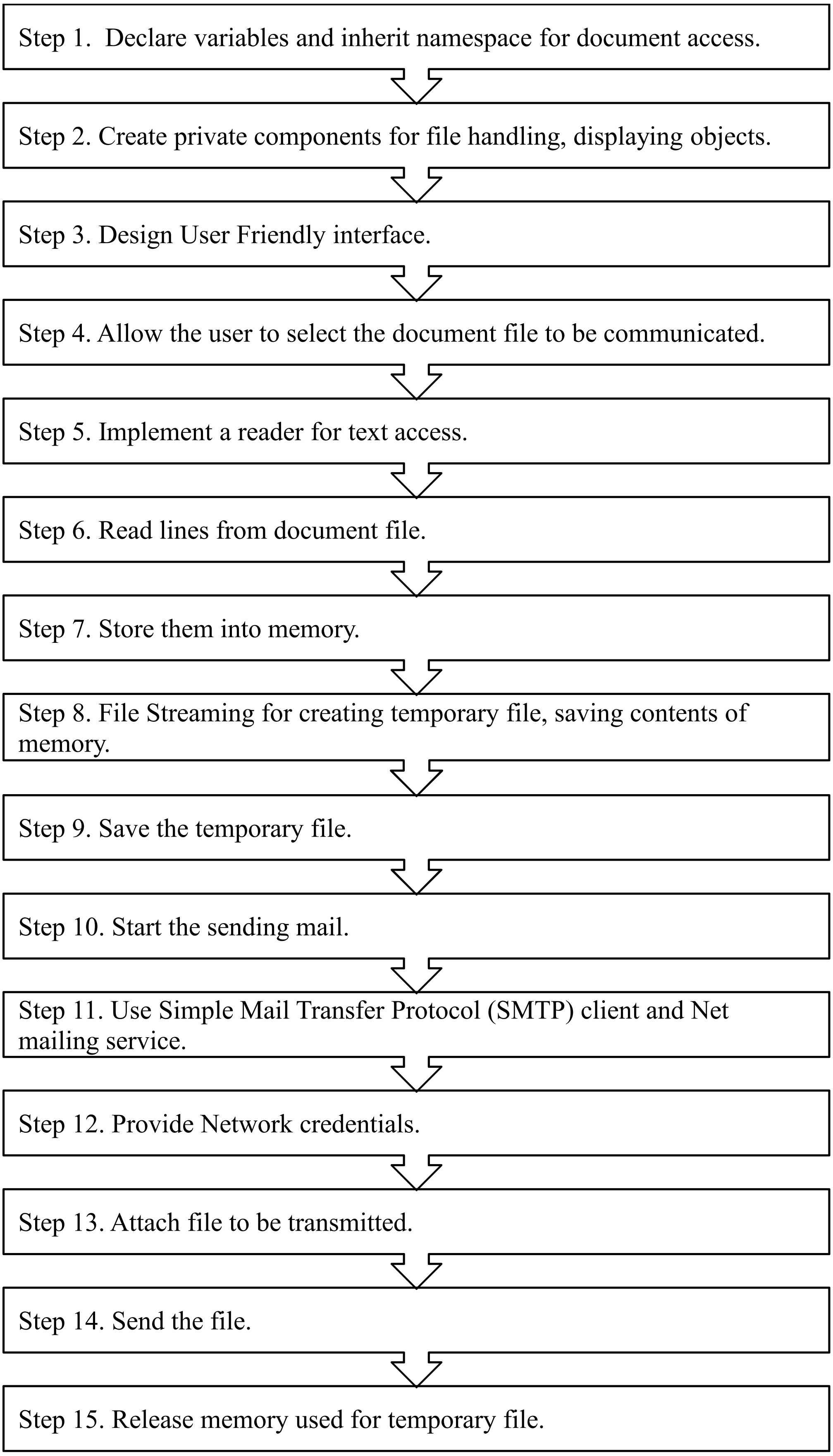
The proposed work introduces a better text communication method which can be used for searching of data [57]. The newspaper industry has been shifting from print media to electronic form. The textual information in the form of news gets saved on the data servers. The proposed work can be configured for news searching applications. BIT-STREAM in this context is used as parts of streams.

3.6 Bit-Streaming for Text Communication

The carried out work proposes a prototype for text extraction and exchanging over the email networks. Embedded feature of character only extraction makes it a unique kind of model. The bit streaming prototype has been proposed for the text communication.

3.7 Algorithm Description

In this proposed work, a model prototype algorithm has been defined. This algorithm is the stepwise execution of the program steps for extracting characters and exchanging the plain text character files. The first step is to assign the memory variables and the document access. The derivation for file handling functions is essential. Basic file handling functions are already developed. This prototype inherits these functions and their features. The program prototype has the user friendly interface for the application. The user has been guided for using the prototype model. This interaction area enables user to do the file selection and text extraction activities. Input is sought by the application in the form of document file having formatting information embedded. User selects the file for exchange. This file is the source file for the proposed application. The advanced code modules for stream reading of text matter are executed for file reading. The characters are read from the document files. These file contents are temporarily stored in the computer's memory. These lines are processed through the character extraction prototype. This prototype is developed with addition of the network protocol for sending file through electronic mail. This process saves time. The credentials are properly configured and set with the email account. The file with plain text characters is shared with the mail server with the receivers account. After the successful execution of the prototype the allocated memory and function inheritors has to be released [58].The steps of algorithm are listed out in the **Flowchart 3.1**:



Step 1. Declare variables and inherit namespace for document access.

Step 2. Create private components for file handling, displaying objects.

Step 3. Design User Friendly interface.

Step 4. Allow the user to select the document file to be communicated.

Step 5. Implement a reader for text access.

Step 6. Read lines from document file.

Step 7. Store them into memory.

Step 8. File Streaming for creating temporary file, saving contents of memory.

Step 9. Save the temporary file.

Step 10. Start the sending mail.

Step 11. Use Simple Mail Transfer Protocol (SMTP) client and Net mailing service.

Step 12. Provide Network credentials.

Step 13. Attach file to be transmitted.

Step 14. Send the file.

Step 15. Release memory used for temporary file.

Flowchart 3.1: Steps of proposed algorithm

3.8 Experimental Setup

The proposed experiments were carried out on HP Pavilion dv2700 Notebook PC using Visual Basic language for code development.

3.9 Results and Discussion

Experimental results show the file size reduction in original file size. A document file with 612 words is processed for BIT-STREAM. The plain text file's original size is 31,744 bytes. After character extraction the file size is reduced 7.719844358 times to the size of 4,112 bytes. Ultimately, the file transfer becomes faster. It also shows the bandwidth saved in number of bytes. The result has been summarized in **Table 3.1**.

Original File	**Original Size in bytes (A)**	**Bit-Stream Size in bytes (C)**	**Bandwidth save in bytes (A-C)**	**Original Size reduction (number of times)**
PlainText.doc (612 words)	31,744	4,112	27,632	7.719
FormatText.doc (612 words)	32,768	4,566	28,202	7.176

Table 3.1: Experimental Results

3.10 Concluding Remarks

The importance of text communication and basics of text communication are discussed in this chapter. The novel concept of character streaming using a bit extraction mechanism has been proposed. An algorithm for the proposed prototype is presented in this chapter. As the journalism is based on the textual communication paradigm, the important aspects, limitations and features of text communication are discussed here. A proposed model and its output analysis has been stated in this chapter.