Bit-Stream Text Communication for Effective Use of Bandwidth and Faster Retrieval

Narangale S. M.
School of Media Studies,
SRTM University,
Nanded-431606, INDIA

G. N. Shinde*
Indira Gandhi College (Sr.), CIDCO,
Nanded-431603, INDIA

**Address for communication*

Abstract - **This paper proposes a unique system for text file communication for effective use of bandwidth and faster retrieval. When document files are exchanged over mail servers, they include formatting information that increases text file size manifold. Bit-stream presents a technique that extracts only characters from formatted document files resulting in lowered message transfer time and file size. The major objectives of this experiment are to maximize:**

- **the reaction speed of information retrieval**
- **the effective use of internet bandwidth and**
- **the use of offsite resources**

and to minimize:

- **the file exchange time**
- **the load on mail servers and**
- **the restrictions of proprietary software use.**

The file size in BIT-STREAM Text communication gets reduced upto 86.55%, which is comparatively higher than any other compression techniques.

Keywords-Text Communication; BIT-STREAM; offshore resources; Compression; character extraction; bandwidth; mail server;

I. INTRODUCTION

Messages are stored in the mail servers. Gmail, Yahoo, Rediffmail, Hotmail typically provide around 1GB storage space per user account by limiting one email attachment to around 20-25MB. Message transfer over internet (online storage, attachment, and share) contains lot of formatting information leading to large file size. Mail servers have limited resources, which include storage systems, retrieval systems, ranking systems, etc. IBM's review claims to create more than 2.5 quintillion bytes of data daily and to meet this criteria any mechanical/ magnetic device require huge resources[1]. Digital storage formats have gained maximum share among all the storage media after 2000. In 2011, 94% of total technological memory (storage) are in digital format [2]. Australian Bureau of Statistics claims: the overall volume of data downloaded increases by 26.0% since June 2011 to 345,518 Terabytes in Australia [3].

E-mail transfer is growing at a tremendous rate. In 1995, 100 billion e-mails were sent annually; in 2002 this number had increased to 5.5 trillion e-mails; in 2010 the amount of emails sent daily is expected to be around 294 billion that means 90 trillion mails are sent per year [4]. About 85% of all email data are due to attachments. People send 20% of their time searching through their emails and files. Stubbing is one way, through which links can be sent to the receiver. These links can contain images, video, formatted text, etc. This way the load on email servers can be reduced.

Documents with formatting like Microsoft's word changes the file format with each release, so, its users are locked into a system that compels them to buy each upgrade whether they want a change or not. MS-Word is a proprietary and secret document format. People may not have the proprietary software that can read secret format documents. A typical one page document can be ten times bigger and slower in Word than in plain text. Many mail servers have limited space in their email accounts.

Speeds of communications channels, both wired and wireless are increasing steadily but not dramatically [5]. No one has configured a cost-effective way to reduce file size, to store emails, access and search them. The present work is an attempt to design an adaptive BIT-STREAM mechanism to perform segmentation of characters from documents containing characters and formatting information. The proposed system is the application of BIT-STREAM mechanism. Characters from document files can be automatically found out by stream readers using C#.Net programming language, and can be separated. With the file streams, the separated characters can be saved into temporary file. This type of 'characters only' file is ultimately small in size, results in minimizing the file transfer time.

The rest of the paper focuses on the work already done in relation with text communication and reduction in file sizes. BIT-STREAM communication and its relation with text

communication are discussed. This novel system can be effectively implemented for saving space, retrieving information faster, and doesn't abide users by the restriction of using proprietary software.

II. Related Work

E-mail transfer is growing at a tremendous rate. In 2010 the amount of emails sent daily was around 294 billion that means 90 trillion mails were sent per year [4]. Out of them a huge number of mails contained files as attachments. People generally use Word files as an attachment for sending official documents, resume, reports, research papers, thesis, etc. This exchange of attachments requires ample amount of storage space, and the need for storage is increasing over time. Compression technique is one of the solutions for reducing file size. The data compression software reduces the size of the data file by a factor of two, or results in a "compression ratio" of 2:1 [6, 7]. There are two types of data compression; lossy and lossless data compression. In lossless data compression, data needs to be uncompressed exactly as it was before compression whereas in Lossy compression, data doesn't have to be stored perfectly. Even though some data is lost, when uncompressed, the data will still be of acceptable quality. Text files are stored using lossless techniques, since losing a single character can be in the worst case make the text dangerously misleading [8].

A Radicati report explains, a typical user sends/receives messages over email in 2009 about 167; in 2010 about 179; in 2011 about 192 and in 2012 about 205. Hence, email servers, who store these mails, need more storage space. Out of the total number of messages in 2009, the messages sent/received with attachments were 37. The average storage per corporate user per day required was 20.3MB. And, this number is continuously increasing: in 2010 it was 22.3MB, in 2011 it was 25MB, in 2012 it is about to reach 27.6MB and in 2013 it is projected to cross 30.8MB [9].

Compression reduces file sizes to a considerable amount. Run length encoding or run length limiting is a simple compression technique. In it, if we have a text file in which the same characters are often repeated one after another. Run length encoding compresses the file by removing redundancy of characters. Run length encoding is actually not meant for compressing text files, because, a text file has meaningful information, and doesn't have a lot of long, repetitive character strings [8].

Another Compression technique is Huffman coding, in which, the characters in a data file are converted to a binary code, where the most common characters in the file have the shortest binary codes, and the least common have the longest [8].

LZ-77 uses pointers to previous words or parts of words in a file to obtain compression. Lempel and Ziv came up with an improved scheme in 1978, appropriately named LZ-78, and it was refined by a Mr. Terry Welch in 1984, making it LZW [10,11]. LZW takes that scheme one step further, actually constructing a "dictionary" of words or parts of words in a message, and then using pointers to the words in the dictionary.

III. BIT-STREAM Communication

BIT-STREAM refers to the retrieval of data from different resources by performing simple operations [12]. Data can be extracted in the form of a stream of characters from document files. BIT-STREAM in this context is used as bits and pieces of the media stream. The proposed system is the application of BIT-STREAM mechanism. Characters from document files are treated as a group in the form of a stream. It can be automatically found out by stream readers using C#.Net programming language, and can be separated.

IV. BIT-STREAM for Text Communication

For this BIT-STREAM program to execute, we inherit the GetDocText.Doc namespace. Then we create private components for container, file handler, display objects, action listeners, etc. User friendly interface is built. At the start event, the dlgOpenFile.ShowDialog will handle the selection of files. The selected file is then checked for its format/ type. If the selected file is of MS-Word document format (*.doc), with formatting information, it is forwarded for character extraction. For reading the text matter from selected files advanced class of StreamReader is used. StreamReader class is a container for a method headline that reads a stream of characters from a file in a block of 255 characters. Figure 1 indicates a sample code for this procedure.

Using the FileStream class a temporary file is created. UTF Encoded bytes are extracted from the document file. Extracted characters are stored in a temporary file. The temporarily created file has a very less size compared to original document file. Net. Mail namespace is necessary for sending mail with attachment through the mail server. It is possible to send the mail from BIT-STREAM desktop application. Mail will be sent by creating SMTP client and providing necessary network credentials to it. It is observed that the mail is sent with attachments.

At the receiver side, the automated control object is able to identify the type of file to be accessed through email. File

handler reads plain text data from attachment, creates a document, saves the text into a document file.

The receiver is free from the restrictions of using proprietary software. This application confirms from the user, whether to use it with proprietary software or not.

V. EXPERIMENTAL SETUP

The proposed system is executed on the HP Pavilion dv2700 Notebook PC with Intel (R) Pentium (R) Dual CPU @ 1.73GHz processor and 2GB RAM and Microsoft Visual Studio 2008 with C#.Net as the development framework. The performance of BIT-STREAM was tested on two doc files; one with plain text characters and other having formatting information.

A. Algorithm

Algorithm: BIT-STREAM Text communication

START:
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. Implement File Streaming for creating temporary file, saving the contents of memory.
Step 9. Save the temporary file.
Step 10. Start the sending mail.
Step 11. Use 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.
STOP.

B. Snapshots

In snapshots, figure 1 is of sample program code. It shows the sample code for accessing the document file contents. Figure 2 is of User Interface form for selecting a document file. The user selects the document file to be communicated over email. Figure 3 is an interface window, using it, users can send an attachment to the email recipient.

```
using System.Text;
using GetDocText.Doc;
StreamReader sr = new StreamReader(dlgOpenFile.FileName);
text = sr.ReadLine();
text = text + sr.ReadLine();
FileStream fs = File.Create(path);
AddText(fs, text);
UTF8Encoding temp = new UTF8Encoding(true);
while (fs.Read(b,0,b.Length) > 0)
{
Console.WriteLine(temp.GetString(b));
}
private static void AddText(FileStream fs, string value)
{
byte[] info = new UTF8Encoding(true).GetBytes(value);
fs.Write(info, 0, info.Length);
}
```

Figure 1:Snapshot of Sample program code.

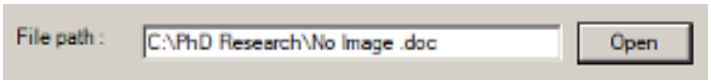

Figure 2:Snapshot of Form for selecting *.doc file.

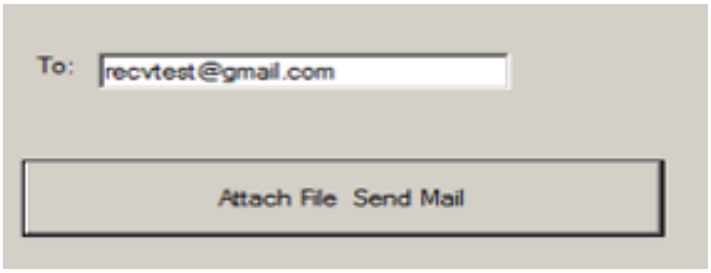

Figure 3: Snapshot of file attachment & sendmail.

VI. RESULT AND DISCUSSION

The results show the file size reduction in original file size. A plain text document file with 612 words of size 31,744 bytes is processed for BIT-STREAM. After RAR compression file size is reduced to 8,541 bytes i.e. 73.09% reduction; after ZIP compression file size is reduced to 8,818 bytes i.e. 72.22%; and, the BIT-STREAM character extraction reduces file size to 4,112 bytes with reduction upto 87.05% showing considerable increase. In case of formatted text document file of 612 words of size 32,768 bytes; the RAR compression reduces file size to 9,098 bytes i.e. 72.23%; ZIP reduces file size to 9,402 bytes i.e. 71.30%; and, the BIT-STREAM reduces file size to 4,566 bytes i.e. 86.06% reduction. Ultimately, the file transfer becomes faster, bandwidth gets saved, retrieval becomes faster. The result is summarized in table below.

Original File	Original Size in bytes	BIT-STREAM created file	RAR Compressed Size in bytes		ZIP Compressed Size in bytes		Bit-Stream Compressed	
			Size in bytes	% Reduction	Size in bytes	% Reduction	Size in bytes	% Reduction
PlainText.doc (612 words)	31,744	PlainTextTest.txt	8,541	73.09%	8,818	72.22%	4,112	**87.05%**
FormatText.doc (612 words)	32,768	FormatTextTest.txt	9,098	72.23%	9,402	71.30%	4,566	**86.06%**

VII. CONCLUSION

In this paper Bit-Stream mode of communication for text messages is presented. From the source of references, this is the first effort towards extraction of characters from formatted text files for exchange through mail servers. As mentioned earlier, most of the people use Microsoft Word as the file transfer format as attachment, which results in consumption of huge size. It becomes necessary to use proprietary software for reading those files. The file size in BIT-STREAM Text communication gets reduced upto 86.55%. Whereas in RAR, file size is reduced upto 73% and using ZIP techniques file size is reduced upto 72%. Thus, the file size in BIT-STREAM Text communication is comparatively more condensed than RAR and ZIP techniques. This model saves space, time for exchanging information, keeps no restriction on proprietary software requirements. The communication technique presented here makes information retrieval fast.

VIII. REFERENCES

[1]. White paper, "What is big data?". In IBM's web page Source from *www*-01.ibm.com/software/data/bigdata/

[2]. Hilbert, Martin, and Priscila López. 2011, The World's Technological Capacity to Store, Communicate, and Compute Information. Science 332(6025): pp60-65

[3]. Volume of Data Downloaded. (2012), Source from http://www.abs.gov.au/ausstats/abs@.nsf/Lookup/8153.0Chapter7Dec%202011

[4]. Tschabitscher, H. (2006). How many e-mails are sent every day? About. April 28, 2009, source from http://email.about.com/od/emailtrivia/f/emails_per_day.htm.

[5]. K. R. Kolhe, P. R. Devale and P. Shrivastava, High Performance Lossless Multimedia Data Compression through Improved Dictionary, International Journal of Computer Applications (0975 – 8887), Vol.10, No.1, Nov. 2010

[6]. Miano J.,Compressed Image File Formats: JPEG, PNG, GIF, XBM, BMP, August 1999

[7]. Khalid Sayood, Ed Fox (Editor), Introduction to Data Compression, March 2000

[8]. Mohammed Al-laham and Ibrahiem M. M. El Emary, Comparative Study between Various Algorithms of Data Compression Techniques, International Journal of Computer Science and Network Security, VOL.7 No.4, April 2007, 281-291.

[9]. Sara Radicati, Masha Khmartseva, "Email Statistics Report 2009-2013", The Radicati Group, Inc, Palo Alto, CA 94301. Source from http://www.radicati.com/wp/wp-content/uploads/2009/05/email-stats-report-exec-summary.pdf

[10]. Fernando C. Pereira (Editor), Touradj Ebrahimi, July 2002

[11]. Bentley, J. L., Sleator, D. D., Tarjan, R. E., and Wei, V. K. 1986. A Locally Adaptive Data Compression Scheme. Commun. ACM 29, 4 (Apr.), 320-330.

[12]. Gabriel Panis, Andreas Hutter, Jorg Heuer, Hermann Hellwagner, Harald Kosch, Christian Timmerer, Sylvain Devillers, Myriam Amielh, Bitstream syntax description: a tool for multimedia resource adaptation within MPEG-21, Signal Processing: Image Communication 18 (2003) 721–747.

Biographies

Mr. Sachin M. Narangale: Is Assistant Professor in the School of Media Studies, Swami Ramanand Teerth Marathwada University since Feb. 2009. He was born on 20th August 1980. He received his B.Sc. in 2001 with Physics, Electronics and Computer Science; and MCA in 2004 from SRTM University, Nanded. He has University Rank third in MCA. He has worked with DSRM Embedded Technologies (Inc. 500 listed, CMM Level 3, US based company) as a Software Engineer around 2 years. His interests include compression, multimedia communication, software development and embedded systems.

Dr. G. N. Shinde: Is Principal of Indira Gandhi College, Nanded, Maharashtra, INDIA. He has received "Ideal State Teacher Award" from Government of Maharashtra, India for 2008-09 and "Best Principal Award" for 2009-2010 from S.R.T.M. University, Nanded, Maharashtra. He has received an M. Sc. & Ph.D. degree from Dr. B. A. M. University, Aurangabad.

He was awarded the Benjonji Jalnawala award for securing highest marks at B.Sc. He is Dean, faculty of Science of S. R. T. M. University, Nanded. Eleven research scholars were awarded a Ph.D. degree under his guidance. He has published more than 75 papers in the International Journals and presented more than 50 papers in International Conferences. He was the Chairperson for F-9 session of "International Conference on Computational and Experimental Science & Engineering" (ICCES2008) which was held at Honolulu, U.S.A & Development and Application of Web 2.0 Technology for Education Purpose session of "International Conference on Multimedia & ICT in Education (m-ICT2009)", April 24, 2009, (Hall 1), Lisbon (Portugal) and "Wave Propagation and Wave Interaction with media" Progress In Electromagnetic Research Symposium (PIERS), March 25, 2010, Session 3A4 (Room D**)** Xian, China. In his account one book is published, which is a reference book for different courses. He is also member of different academic & professional bodies such as IAENG (Hon Kong), ANAS (Jordan). He is in reviewer panel for different Journals such as IEEE (Transactions on Neural Networks), International Journal of Physical Sciences (U.S.A.), Journal of Electromagnetic Waves and Applications (JEMWA, U.S.A.). His abroad Visit includes the U.S.A., Thailand, Portugal, Germany, Swisszerland, Italy, Vatican City, Monaco, France, Maldives, Sri Lanka, U. K., Scotland and China. He was Chairman of Grievances Committee and member of Management Council & is member of Academic Council and Senate of S.R.T.M. University, Nanded, INDIA.