NOVEL APPROACH FOR TEXT COMPRESSION

Miss. Pallavi Pralhad Surwade[1],
Prof. Vijay B. More[2]

[1] *M.E. Student, Department of Computer Engineering, Savitribai Phule University, Nasik, India*
[2] *Assistant professor, Department of Computer Engineering, Savitribai Phule University, Nasik, India*

ABSTRACT

Generally, textual data sets are represented by using different models. But, sometimes it does not capture the text arrangement as it is. Vector space model is also recognized as the bag of word model. To represent textual document using vector space model is based on most text mining methods. This model cannot maintain the text structure as it is. Compression distances are the most widely used technique for the purpose of knowledge discovery and mining of data also to improve the performance metric. Compression distance technique is for measuring the similarity between two documents. By applying Distortion technique which is for purpose of destroys the text structure. A distortion technique removes nonrelevant words as well as maintains the text structure. The NCD i.e. normalized compression distance technique gives the structural similarity between text documents. By applying a detailed word removal method clustering accuracy can be enhanced. In proposed system, JPG encoding is the technique that converts text into pixels of an image. The technique to convert text into pixels of an image. JPG encoding technique wherein an image pixel contains the values of RGB. So, it greatly reduces the size of the document. Each pixel value assigns an equivalent English letter of the same byte value where each pixel contains 3 letters. To reduce the size of text documents, For getting the lossless compression, Encodes the text document in term of image format.

Keyword: - *Data compression, Data distortion, Normalized compression distance, Text mining, Text representation.*

---

1. INTRODUCTION

Vector Space Model or Bag-of-word model is an arithmetical model mostly designed for representation of text. A usual method of explanation of relations among words (text structure) is related with compression distances. Though, text formation gives significant information on a data or text, but unmodified or unchanged texts have words which are not relevant in nature that can make the comparison of texts complicated [1] [2]. Apply distortion technique by means of helping the compressor get additional validated similarities in a compression-based clustering scenario; this distortion technique eliminates the information which is not relevant also preserves relevant or significant information and also the structure of text. This is made by eliminating the nearly all common or repeated words in the English language from the documents, and replaces each of their characters with an asterisk. These distances give a similarity measure between two objects using data compression. Provide a measure of the similarity between two texts from texts themselves. To determine the data sets structural characteristics, the NCD-based clustering can be used, also examines how destroying text structure affects the NCD. A compressed or reduced file is any le that includes one or more les or directory that is smaller in size than their original size of the file. So that downloading gets faster, easier and allows additional data to be store on a removable media. Compression of data is advantageous.

Decompression is thus important for compressed data as all compressed data needs to be decompressed. The application needed for decompression largely depends on how the data was compressed in the first place. There are different techniques and algorithms available for decompression of data. So, the proposed system is consists of

the encoding as well as the decoding technique, so it gives us the lossless data. While decompression of the data, in the lossless compression, original data is gained with no loss. Data decompression facilitates in eliminating the complication or the problems added by data compression. So, the proposed system is implemented to greatly reduce the size of the document, as compared to other techniques. The NCD that is the normalized compression distances gives the similarity between two words. The use of the NCD based clustering technique as one tool compute or measures the performance of the NCD, allows analysis of how the information contained in the texts develops as words are detached from the texts. The management of textual data, the idea of context is very useful because it is powerfully bound to texts. The text has the consistent structure, applying the thought of context to text management occurs naturally. Since, the parameter-free nature, wide applicability of the NCD is applied to many research areas and their leading efficiency. Among others, they have applied to document retrieval, document clustering, security of computer system, music classification, data mining, plagiarism detection, software engineering, bioinformatics, chemistry, medicine or even art [1], [5]. Normalized compression distances technique also applied to music classification. In proposed system, encoding as well as decoding technique is proposed.
JPG encoding is the technique that converts text into pixels of an image. The technique converts text into pixels of an image. JPG encoding technique wherein an image pixel contains the values of RGB. So, it greatly reduces the size of the document. Each pixel value assigns an equivalent English letter of the same byte value where each pixel contains 3 letters. To reduce the size of text documents also for getting the lossless compression and Encodes the text document in term of image format All musical pieces are same to each other; also some are the largest part of other pieces. For well-organized music information recovery or retrieval systems these similarities are also important. Now, the huge amount of digitized music is accessible on the internet is growing, on commercial sites and in public domain. Websites giving musical content are of the form such as MP3, MIDI or other. So, with the musical genres and subgenres, putting similar pieces together, classify their files. Organizations principle is to allow users to find the pieces of music which are well-known to them and also give those recommendations and advice. Compression distances are applied in many research areas like question answering systems, data mining [2], plagiarism detection [4], and philology, information retrieval, and text categorization. Data compression technique gives a similarity measure between two objects. It means they calculate the relationship between two texts from texts themselves. Texts can be represented can be represented directly, It doesn't always required any model. Examine several compression algorithms responsible for capturing text structure. Particularly, PPMD, BZIP2, and LZMA are examining. They are of dissimilar compressors families: PPMD is an adaptive arithmetical compressor, BZIP2 is a block-sorting compressor and LZMA is one dictionary compressor. By using a compression algorithm that permits its order to be changed carried out an experimental analysis which is based on this.

2. RELATED WORK

Ana Granados, G.Salton, et al. [1], [2], introduced Vector space model (VSM) which is also recognized as the bag-of-words model is an algebraic model. Here every text document is formed as a linear vector that represents the words occurrence which is not dependant in the text set. Many text clustering methods are supported to the representation of texts using the model i.e. Vector Space Model (VSM). VSM is successfully and mostly useful in several research areas. Representation of text documents by this model does not maintain word arrangement whereas word order is enormously important while representation. There is need to improve this Vector Space Model. Also need to represent text document in a specific manner, by applying distortion technique. Better clustering results can be obtained by maintaining contextual information. Because of improvement in the representation of documents that will allows increasing the accuracy of the results obtained when searching documents. Here, they have applied NCD to text clustering. Also analyzes how different compression algorithms capture text structure. So, here they have used three compression algorithms to perform the NCD-driven text clustering [6].

G. Cao, J.-Y. Nie, and J. Bai [1], [10] initiated technique for information retrieval, it is the important concept. For the information retrieval proposes approach for language modeling for information retrieval. Goal here is to construct a language model in which a variety of relationships of words are integrated. Relationships between words come from 2 sources: from co-occurrences o and the second is from WordNet. The result gives the important improvements. These results obviously show the advantage into language models of incorporating word relationships.

KostadinKoroutchev, and Francisco e BorjaRodrguez [1] [6], Methodology suggested emphasizes on dealing with texts using compression distances. More specifically, it understands the nature of the compression distances and nature of texts. Nowadays, most of the information stored electronically is stored in text form. In fact, if we think of the time that we spend every day reading emails, news, articles or reports, that most of the information that we use every day is text. This fact makes methods that deal with texts really interesting. Three compression

algorithms are used here to calculate compression distances. Each algorithm is of special family of compressors: PPMZ, BZIP2 and LZMA. PPMZ is an arithmetical data compression algorithm and is based on prediction. Compressor BZIP2 is based on the Human codes, the Move-To-Front transform, Burrows-Wheeler Transform and Run Length Encoding. Compressor LZMA is a Lempel-Ziv-Markov chain algorithm.

M. Burrows and D. Wheeler,[14] , A. Lempel [16], implemented one of the lossless data compression algorithm, that is block-sorting. Also, compared the performance of implemented block sorting algorithm with the available data compressors algorithms. This algorithm takes input as a block of text one at a time by applying a reversible transformation to each block to form another new block that includes the similar characters. This algorithm greatly achieves speed with the comparison of the algorithms based on the methods of Lempel and Ziv, but also achieves compression closed with the statistical modeling techniques. The input blocks size has to be big enough such as a few kilobytes for the purpose of achieving the good and lossless compression [16]. A. Hotho, S. Staab, and G. Stumme [13], also clustering of text is one of the important terms for providing in text document clustering is important in providing browsing and perceptive navigation by managing huge sets of documents into a small number of significant clusters. For these clustering methods the bag of words representation is used.
But, frequently it does not pay attention on relationships among terms. So to deal with such type of problem, core ontologies are integrated in the process of clustering of text documents.

R. Cilibrasi and P. Vitanyi [7], presented a novel method based on compression for clustering. The NCD is not applied for any particular application area. To conclude the structural characteristics of data sets the clustering based on NCD can be used. By using the compressor PPMZ the NCD distance matrix was computed. PPM i.e. Prediction with partial string matching uses complicated data structures also it gains the good performance of some actual compressor even as it is typically the slowest and mainly memory concentrated. The clustering used is dendrograms based which is hierarchical clustering. The results of the clustering algorithm is shown by the CompLearn Toolkit, and the output represented as a dendrogram.

R. Chau, A. C. Tsoi [3], have introduced a largely automatic approach to extract the underlying concepts of a document, and also to shows their relationships accordingly in un-directed graphs, i.e. the links between the concepts are having no directions. They call proposed approach a Concept Link graph approach, when the concepts extracted or drawn out through this largely automatic means might be very different to the concepts extracted manually in the concept graph approach. Though sometimes it may contain small phrase, compound words, or are single word in nature, this nomenclature reflects largely the fact that most of the concepts extracted using method proposed by them. Also, they can automatically extract links among these concepts. The Concept Link graph is an approach which is inspired by the bag of words [3] approach except that representing the underlying concepts in the document; also they have modified some of the underlying procedures so that words can be extracted.

A. Granados, M. Cebrian, D. Camacho, and F. de BorjaRodrguez [4], [11], suggested distortion technique has the idea of finding out the data set's organization way. Here, emphasis is on knowledge of the compression distances with a new evaluation which of the in influence belonging of some type of information distortion on compression-based text clustering. [11]Distortion technique applied on unmodied text which contains non relevant information which makes comparison of text difficult. Showing of relevant or useful information from the texts is by eliminating the most occurring words from the texts in the English language, and replaces every character by an asterisks of the detached words[5]. By removing words, the complicatedness of a document is decreases slowly that facilitates the compression based text clustering and progresses its correctness. The text clustering on non-distorted text can be enhanced by annealing text distortion. The results expected to be consistent for several data sets, and compression algorithms which is of different compression families: Block-Sorting, Lempel Ziv and Statistical.

Manuel Cebria n, David Camacho ET. al. [5] Given a step to know or understand the Compression distances also performs an evaluation of the influence of numerous types of word elimination on compression-based text clustering. For implementation of the clustering algorithm they have used the CompLearn Toolkit. So, the several distortion techniques that consists of the loss of information is to be studied.

A. Granados [6], R. Martinez, D. Camacho, and F. B. Rodriguez [15], gives emphasis to know the structure of texts also the nature of compression distances. So, here is that the information in the texts can get changed. The compressor can better capture their structure, and therefore, the obtained NCD-based clustering results can be improved. But sometime unmodified texts contain words which are not relevant but, so, the evaluation
of the texts becomes difficult. A distortion method is applied that helps the compressor to get more reliable similarities in a compression-based clustering. Distortion technique eliminates irrelevant information also preserves text structure and relevant information. The idea is to change the representation of the texts without any loss of important information so; this gives new representation more suitable for compressors for getting similarities between the texts. Also more concern is on the improvement of the correctness of similarity distances from

the compression distances family, the normalized compression distance (NCD), is used to compute similarities between documents [15]. So, the combination of the document distortion technique with the document segmentation strategy helps to improve the correctness of the NCD when applied to compute similarities between the documents [15].

R. Cilibrasi, Rudi Cilibrasi, R. de Wolf ET. [7], [9], offered a method for music classication, which is depends only on compression of strings that gives the music pieces. Computed the detachment between all pairs of pieces, ensuing in a distance matrix of pair wise NCDs. Here, for the classification of pieces of music compression-based method is applied. In the Musical Instrument Digital Interface, on sets of classical pieces those occurred mostly, performed a variety of experiments. Results are of the form of distance matrix of pair wise NCDs, computed the distances in all pairs of pieces. This result shows that in a hierarchical tree contains pieces which are consistent with the computed distances.

3. SYSTEM ARCHITECTURE

As below figure 1 show, Compression distance techniques, it helps us to cluster and retrieve documents. Compression distances use compression algorithms to calculate the similarity between two objects. The distortion technique that direct to an improvement in the non distorted clustering, results consists in eliminating the most common words of the English language from the documents, substitutes their characters with asterisks.
This strategy allows preservation of the text structure despite the word removal. Then, NCD i.e. Normalized Compression distances is use to measure the structural similarity between textual documents in, and between XML documents in. NCD is to determine the relationship between texts, some Compression algorithms used to compute the NCD. one of the technique that is JPG encoding technique included here. JPG encoding technique converts text into pixels of an image. JPG encoding technique wherein an image pixel contains the values of RGB and each value accepts 0 to 255. Each pixel value assigns an equivalent English letter of the same byte value where each pixel contains 3 letters.

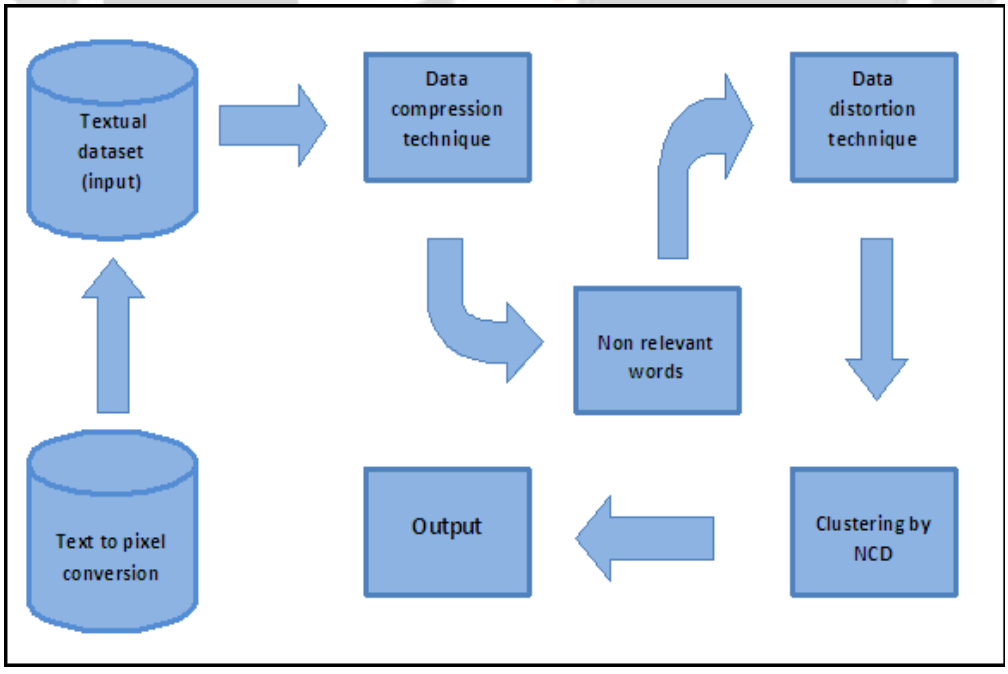

Fig.1 System architecture diagram

4. ALGORITHMIC STRATEGY

4.1. Compression algorithm

1. Compressor LZMA, that is, it is a Lempel-Ziv-Markov chain algorithm.
2. Compressor PPMZ is an arithmetical data compression algorithm based on prediction.
3. Compressor Block-sorting i.e.BZIP2 is a compressor based on the Burrows-Wheeler Transform, Huffman codes, Run Length Encoding, the Move-To-Front transform. Compression algorithms for NCD-driven text clustering.

(A) Dictionary Methods: LZMA

Step- by – step execution-

Step I- First here take out the smallest substring that is not found in the residual string which is uncompressed.
Step II- Save that substring as a new entry in the dictionary and allocate it an index or directory value.
Step III- Substring is substituted with the index originated in the dictionary.
Step IV- Keep last character as it of the substring is and also insert the index into the compressed string.

(B) Block-Based Methods: BZIP2
Compresses data using Run Length Encoding (RLE).

4.2. Clustering by NCD Algorithm

The NCD-based clustering permitted to examine how the analyze distortion techniques demolishes the text structure. Normalized Compression Distance NCD, having mathematical formulation is as follows:

$$NCD(x,y) = \frac{\max\{C(xy)-C(x),C(yx)-C(y)\}}{\max\{C(x),C(y)\}}.$$

Where:
C is a compression algorithm.
C(x) is the size of the compressed version of x.
C (y) is the size of the compressed version of y.
C (xy) is the compressed size of the concatenation of x and y.
C (yx) is the compressed size of the concatenation of y and x.

4.3. JPG encoding algorithm

Step- by – step execution.
Step I- Conversion of String into bytes.
Step II- Creation of bitmap image.
Step III- Assign three characters to RGB value.
Step IV- Stores an image.

Step by step Execution-

(A) Encoding -
1. Assume that a string with 8 characters (example - abcdefgh). After putting this on a byte array, It will gives byte array with the size of 8. The following diagram shows how these ASCII characters can store in an array.
| 97 | 98 | 99 | 100 | 101 | 102 | 103 | 104 |

2. One character will need 8 bits if the characters are signifies with ASCII. A set of 8 bits can signify 256 different characters. But if assume the current application, a simple SMS might be take only near 26 different characters. Therefore it is enough to have 5 bit encoding which can give up to 32 different characters to represent. To transfer to 5 bit, substitute new values to the above characters.
| a = 1 | b=2 | c=3 | d=4 | e=5 | f=6 | g=7 | h=8 |

3. At the new byte array, it will appear like the following (the values of characters are in binary representation).
00000001|00000010|00000011|00000100|00000101|00000110|00000111|00001000|

4. For storing the 8 characters use 8 bytes. After that, from the position of 3rd bit from the left side cut each byte and take out the 5 least significant bits. The result will be as follows-
|00001|00010|00011|00100|00101|00110|00111|01000|

5. In an array of bytes reorganize these bits as follows:
|**00001**000|10**00011**0|0100**0010|1**0011**00|111**01000|
This gives the reduced 8 bytes to 5 bytes. The subsequent section shows how these 5 bytes modify to the 8 bytes and get the original information.

(B)Decoding -
1. Every byte should be represented in to binary, when an array of bytes is given. Then all the 1s and 0s should be set as their index and then can be divide to the sets of five bits. After dividing, it will be as follows:
|**00001** 000|10 **00011** 0|0100 **0010|1** 00110 **00|111** 01000|

2. These sets can be transformed to decimals and these values show the characters that we have encoded.

|**00001** = 1(a)

000|10 = 2 (b)

00011 = 3(c)

0|0100 = 4 (d)

0010|1 = 5 (e)

00110 = 6 (f)

00|111 = 7 (g)

01000| = 8 (h), Then the information can be decoded as "abcdefgh".

5. MATHEMATICAL MODEL

Design:
Let S be the system such that,
S = {s, e, X, Y, A, E | Φ}
s -> Initial State
e -> End State
X -> Input = Textual dataset
Y -> Output = Compressed sized document
A -> Algorithms:

I. Gstream algorithm
II. Normalized Compression Distance(NCD)
III. LZMA algorithm
IV JPG encoding algorithm

E -> Entities
E = User

6. ANALYSIS OF RESULTS

Experimental Setup:
The proposed system is tested on different size of textual datasets or documents. The system works by taking input as one of the selected datasets. The selected dataset is processed under JPG encoding compression algorithm the results obtained from this algorithm greatly compresses the size of text document. The text or any data will be converted into pixels of an image. An image pixel contains the four values RGB and each value accepts 0 to 255. Each pixel value will be assigned an equivalent English letter of the same byte value where each pixel contains 3 letters.
2. Dataset used:
The dataset contains various text files. This files can be the human generated files.

1. Compression on different file size by different algorithms

Size of file in bytes	Gstream	LZMA	Image
204360	12120	36625	2948
408722	23872	81283	3142
614184	35908	120630	3397
818566	47776	161264	3586
1023993	61484	211454	2934

Table 1. Table of compression sizes by different algorithm

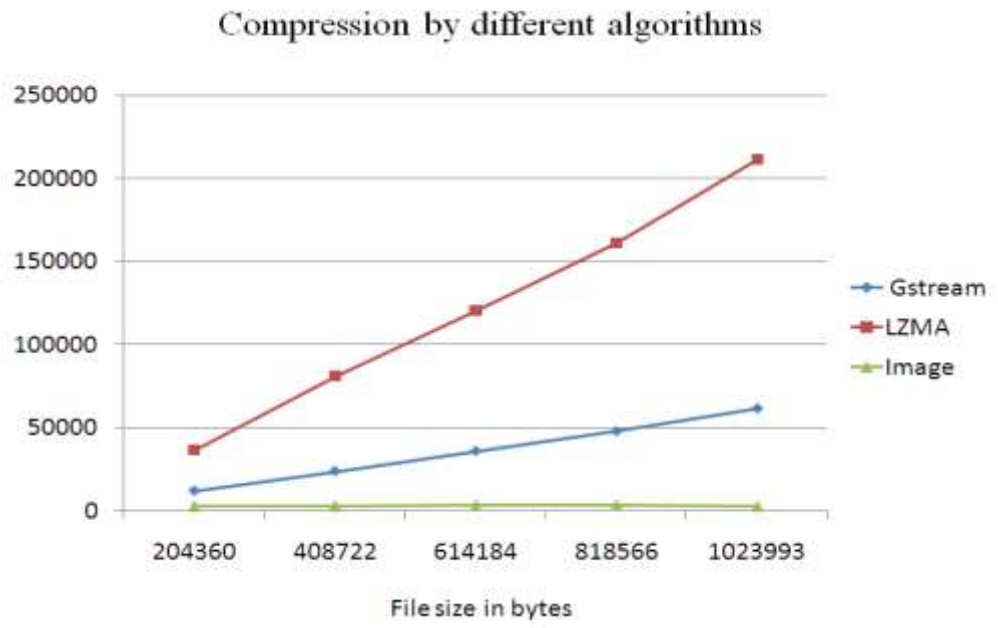

Fig.2 Compression on different file size by different algorithms

The above figure 2 shows the comparison of compression on the different sizes of the various sizes of the files. 'X' axis shows the size of files in bytes. Different algorithms used here, such as the compression by the Gstream algorithm, compression by the LZMA algorithm, as well as the compression by the Image.
So, by the above graph it shows that, Compression by the Image, reduces file size greatly.

2. Decompression on different file size by different algorithms

Size is file in bytes	Gstream	LZMA	Image
204360	204360	204360	204360
408722	408722	410700	408722
614184	614184	615627	614184
818566	818566	820587	818566
1023993	1023993	1026675	1023993

Table 2. Table of decompression sizes by different algorithm

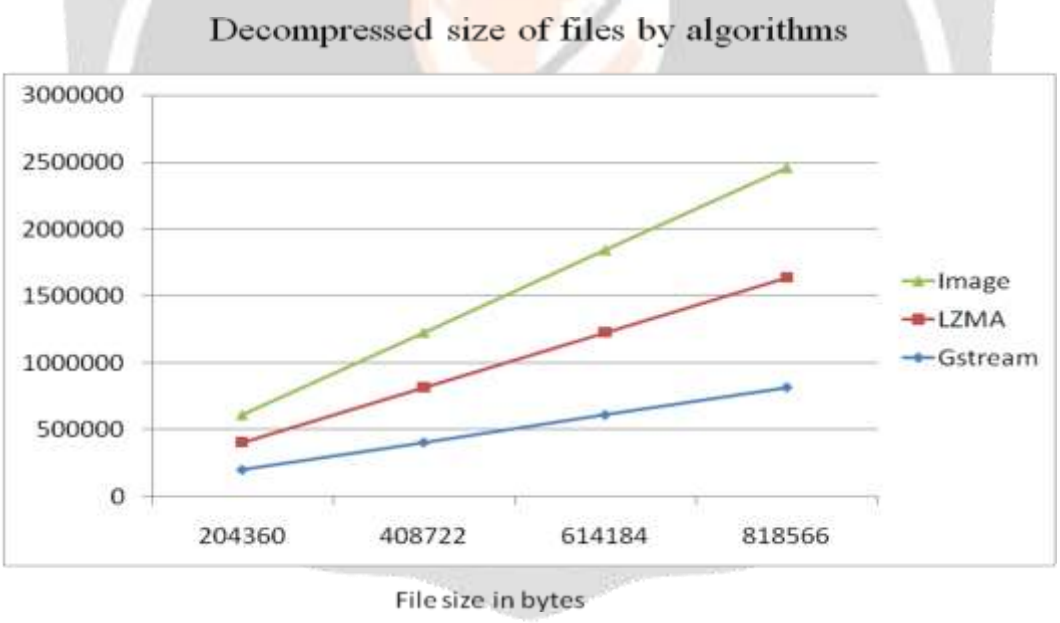

Fig.3 Decompression on different file size by different algorithms

The above figure 3 shows the comparison of deccompression on the different sizes of the various sizes of the files. 'X' axis shows the size of files in bytes. Different algorithms used here, such as the compression by the Gstream algorithm, compression by the LZMA algorithm, as well as the compression by the Image.
Decompression by Image and Decompression by Gstream shows almost same result of file decompression. It means, it preserves text as it it. It gives lossless data as compared with the other algorithm.

3. Compression time on different file size by different algorithms

Size is file in bytes	Gstream	LZMA	Image
204360	49	321	443
408722	42	464	592
614184	56	729	996
818566	52	1036	969
1023993	102	1405	1434

Table 3. Table of compression time by different algorithm

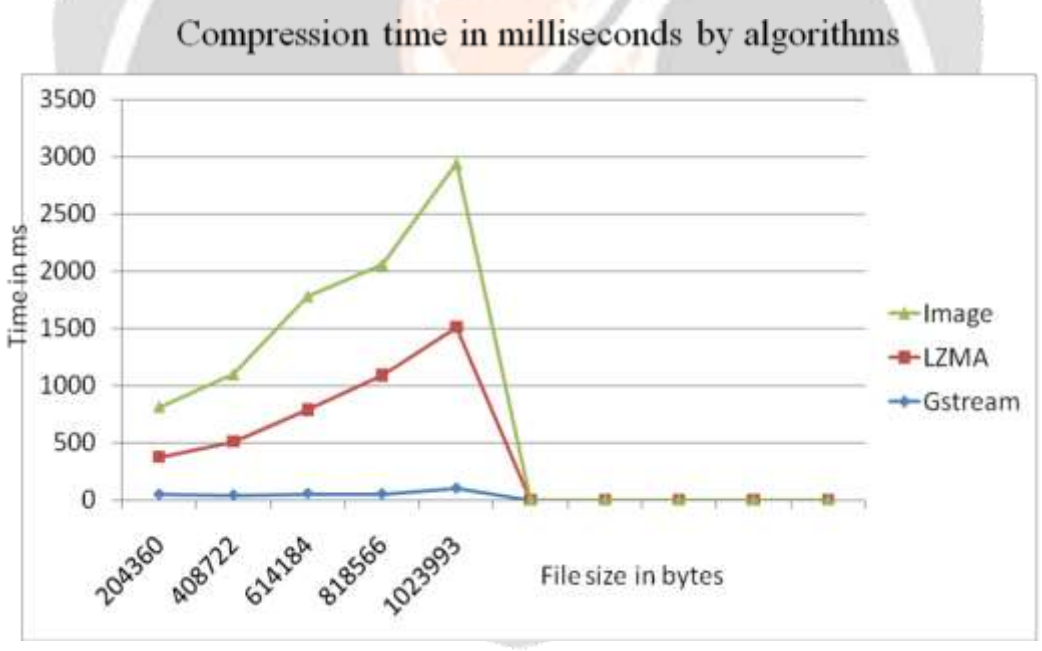

Fig.4 Compression time on different file size by different algorithms

The figure 4 and 5 shows the time taken by the different algorithm for the compression and deccompression on the different sizes of the various sizes of the files. 'X' axis shows the size of files in bytes, And 'Y' axis reprents the time in milliseconds. Different algorithms used here, such as the compression by the Gstream algorithm, compression by the LZMA algorithm, as well as the compression by the Image.

4. Decompression time on different file size by different algorithms

Size is file in bytes	Gstream	LZMA	Image
204360	11	303	15
408722	21	542	33
614184	29	801	38
818566	31	983	48
1023993	82	1331	46

Table 4. Table of decompression time by different algorithm

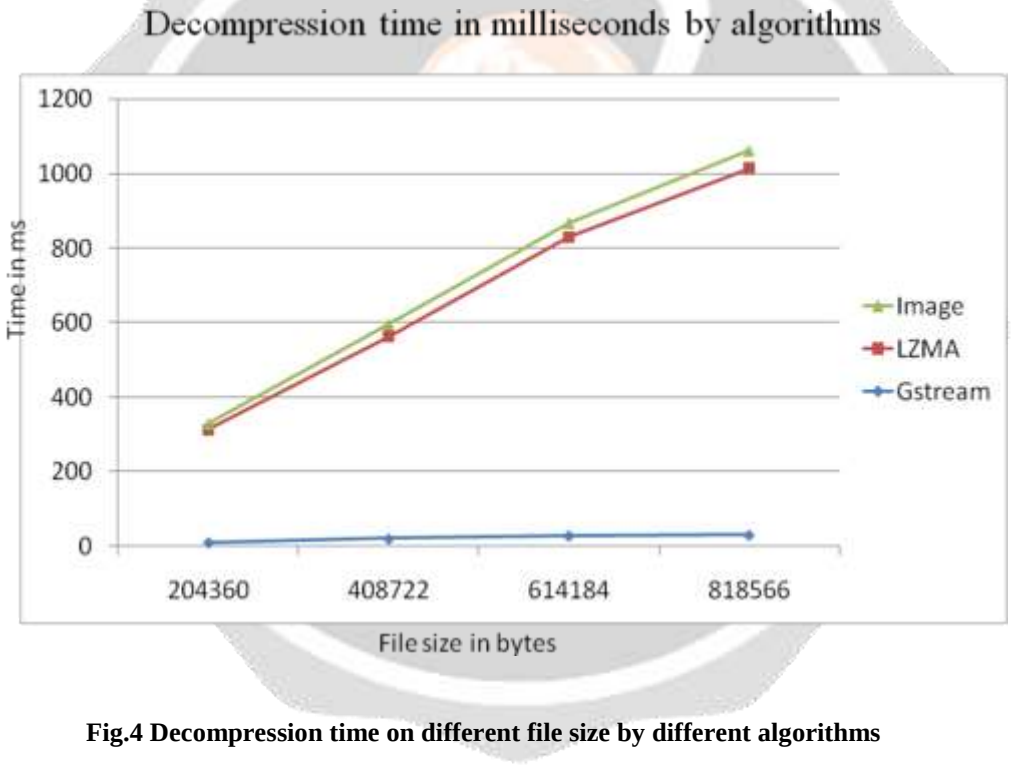

Fig.4 Decompression time on different file size by different algorithms

6. CONCLUSIONS

Reduced sized document is obtained by compression algorithms. NCD is used to measure the similarity between texts, several compression algorithms used to compute the NCD. In proposed system, JPG encoding algorithm will reduce the size of text document. This algorithm encodes the text document in terms of images. Means, converts text into pixels of an image.

Also the system shows the comparison between different algorithms such as the Gstream algorithm, LZMA i.e. Limpel ziv markov chain algorithm and JPG encoding algorithm. Compression as well as decompression by this system done effectively, as compared with another algorithms.

7. REFERENCES

[1] Ana Granados, Kostadin Koroutchev, and Francisco de Borja Rodrguez, Discovering Data Set Nature through Algorithmi Clustering Based on String Compression. IEEE transaction on knowledge and data engineering, vol.27,No.3,March2015.

[2] G. Salton, A.Wong, and C.-S. Yang, A vector space model for automatic indexing, Commun. ACM, vol. 18, no. 11, pp. 613620, 1975. [3] R. Chau, A. C. Tsoi, M. Hagenbuchner, and V. C. S. Lee, A conceptlink graph for text structure mining, in Proc. 32nd Australasian Conf. Comput. Sci., 2009, vol. 91, pp. 141150.

[4] R. J. Mooney and R. Bunescu, Mining knowledge from text using information extraction, ACM SIGKDD Explorations Newslett Natural Language Process. Text Mining, vol. 7, no. 1,

[5] A. Granados, M. Cebrian, D. Camacho, and F. de Borja Rodriguez, Reducing the loss of information through annealing text distortion, IEEE Trans. Knowl. Data Eng., vol. 23, no. 7, pp. 10901102, July 2011.

[6] A. Granados, Analysis and study on text representation to improve the accuracy of the normalized compression distance, AI Commun., vol. 25, no. 4, pp. 381384, 2012.

[7] R. Cilibrasi and P. Vitanyi, Clustering by compression, IEEE Trans. Inf. Theory, vol. 51, no. 4, pp. 15231545, Apr. 2005.

[8] M. Li, X. Chen, X. Li, B. Ma, and P. Vitanyi, The similarity metric, IEEE Trans. Inf. Theory, vol. 50, no. 12, pp. 32503264, Dec. 2004.

[9] R. Cilibrasi, P. Vitanyi, and R. de Wolf, Algorithmic clustering of music based on string compression, Comput. Music J., vol. 28, no. 4, pp. 4967, 2004.

[10] G. Cao, J.-Y. Nie, and J. Bai, Integrating word relationships into language models, in Proc. 28th Annu. Int. ACM SIGIR Conf. Res. Develop. Inf. Retrieval, 2005, pp. 298305

[11] A. Granados, D. Camacho, and F. de Borja Rodriguez, Is the contextual information relevant in text clustering by compression? Expert Syst. Appl., vol. 39, no. 10, pp. 85378546, 2012.

[12] A. Hotho, S. Staab, and G. Stumme, Ontologies improve text document clustering, in Proc. IEEE 3rd Int. Conf. Data Mining, 2003, pp. 541544.

[13] M. Burrows and D. Wheeler, A block-sorting lossless data compression algorithm, Digital SRC Research Report, Tech. Rep. 124, 1994.

[14] A. Granados, R. Martinez, D. Camacho, and F. B. Rodriguez, Improving the accuracy of the normalized compression distance combining document segmentation and document distortion, Knowl. Inf. Syst., vol. 41, no. 1, pp. 223245, 2014.

[15] J. Ziv and A. Lempel, Compression of individual sequences by variable rate coding, IEEE Trans. Inf. Theory, vol. 24, no. 5, pp. 530536, Sept. 1978.