

Final Report for Period: 08/2002 - 07/2005**Submitted on:** 09/22/2005**Principal Investigator:** Mukherjee, Amar .**Award ID:** 0207819**Organization:** U of Central Florida**Title:**

Collaborative: Compressed Domain Search for Text and Images by Sorted Contexts

Project Participants**Senior Personnel****Name:** Mukherjee, Amar**Worked for more than 160 Hours:** Yes**Contribution to Project:****Post-doc****Graduate Student****Name:** Zhang, Nan**Worked for more than 160 Hours:** Yes**Contribution to Project:**

Mr. Zhang is a Ph. D. student working on this project. His primary reserach is concerned with text compression and compressed domain pattern matching for text. Mr. Zhang is working under Prof. Mukherjee's supervision.

Name: Satya, Ravi**Worked for more than 160 Hours:** Yes**Contribution to Project:**

Mr. Ravi Vijaya Satya is working in data compression problems collaborating with Dr. Mukherjee and other graduate students. Mr. Ravi Vijaya Satya is also working on related area of DNA compression and bioinformatics. Mr. Satya is working under Prof. Mukherjee's supervision.

Name: Tao, Tao**Worked for more than 160 Hours:** No**Contribution to Project:**

Mr. Tao Tao is working on compressed domain pattern matching for image problems. He is a Ph. D. student. . Mr. Tao is working under Prof. Mukherjee's supervision.

Name: Sun, Weifeng**Worked for more than 160 Hours:** Yes**Contribution to Project:**

Mr. Weifeng Sun is a prospective Ph. D. student and is working on lossles text compression problems. Mr. Sun is working under Prof. Mukherjee's supervision.

Undergraduate Student**Technician, Programmer****Other Participant****Research Experience for Undergraduates****Organizational Partners**

West Virginia University

Prof. Don Adjeroh, Project Co-PI, West Virginia University

Title: Collaborative Research: Compressed Domain Search for Text and Images by Sorted Contexts

Award #: IIS-0228370 @ West Virginia University.

Other Collaborators or Contacts

We have been collaborating with a well-known researcher in the data compression field: Professor Tim Bell of Computer Science Department, University of Canterbury, New Zealand. Tim Bell is one of the co-Principal Investigators of the project although he has been listed as a Senior Personnel on the budget page of the proposal for technical reasons. His two students Matt Powell and Andrew Firth have contributed to the project. NSF does not fund their activities; they were supported by the resources available to them from the University of Canterbury, New Zealand. We have worked on several joint papers on compressed domain pattern matching.

Activities and Findings**Research and Education Activities:**

Project Summary: Goals and Objectives

We had three major objectives of the proposal

1. To provide methodologies for searching directly on compressed text and images, when the compression is based on the family of compression algorithms that depend on sorted contexts. Such algorithms should perform the search with minimal or no decompression of the compressed database text or image. The query text pattern or query image may or may not be compressed. With respect to compressed domain string matching problem, we considered exact, mismatch/similarity, and/or edit distance computation problems. Further, starting with ideas from text compression by sorted contexts, we developed image compression schemes.
2. To develop search-aware compression schemes that will support compressed-domain search directly on the compressed data, with minimal or no decompression of the compressed database text or image.
3. To develop software tools and create an integrated global compression/search utility website for lossless compression of text and images and for efficient search of large collections directly in their compressed form.

Activities

Research and Education Activities:

1. We developed compressed pattern matching algorithms that achieve the complexity bounds of linear exact pattern matching algorithms for text compressed with the sorted context approach generated by the Burrows-Wheeler transform. The algorithms are based on Boyer-Moore exact pattern matching algorithm, binary search, and q-gram filtration which make use of the sorted contexts generated by the Burrows-Wheeler transform.
2. We studied methods to extend the exact pattern matching algorithms to the problem of compressed domain k-mismatch problem.
3. We completed a detailed comparative evaluation of BWT-based methods of text pattern matching, with complete implementation of the different proposed methods.
4. For image compression, we studied pattern matching algorithm in LZW compressed file and two-dimensional pattern matching algorithms on images compressed by JPEG-LS compression algorithm.
5. We have developed a new method for the encoding stage in prediction-based image compression systems. The approach is based on the BWT transform.

Distribution of Activities

The work on pattern matching on BWT encoded text (items 1, 2, and 3 above) was performed jointly by the two PI's in collaboration with Professor Tim Bell of University of Canterbury, New Zealand. The work on LZW-based pattern matching was performed mainly at the University of Central Florida, by Professor Amar Mukherjee and his students (item 4). The work on BWT-based image compression was performed mainly at West Virginia University by Professor Adjeroh and his students (item 5).

Findings:

The major findings for this research project are as follows.

1. We developed two techniques for searching BWT transformed text using Boyer-Moore algorithm and binary search. The first technique applies the Boyer-Moore algorithm on characters that match while they are decompressed and skips the part of the compressed file that cannot possibly lead to any match. The second technique is based on the observation that the BWT transform contains a sorted list of all substrings of the original string, which can be exploited for rapid searching using a variant of binary search. The decoder only has limited information about the sorted context, but fast q-gram (context) generation and matching algorithms have been developed to exploit this with the help of auxiliary index arrays built in linear time. The algorithm (we call it the QGRAM algorithm) first computes the index arrays for the correspondence between F , the first column of the sorted cyclic matrix in BWT and text T . All exact matches are grouped in consecutive rows of the sorted matrix, which makes the binary searches on F and/or q-grams of the matrix very fast. We also developed a new algorithm, called QGREP, which improved on the sublinear search time of the binary search and QGRAM algorithms. Both techniques are faster than decompress-then-search approach for small number of queries, and binary search is even faster for large number of queries. We provided a more rigorous complexity analysis of the performance of the proposed methods. Also, a detailed empirical comparison of the performance of the proposed methods, with other proposed methods was performed.

2. Algorithms are proposed that solve the k -mismatch problem in worst case time in $O(\min\{m(m-k)A^{\text{power } k}\log(u/A), \mu \log(u/A)\})$ where u is the size of the text, m is the size of the pattern, and A is the size of the symbol alphabet. All exact matches are grouped in consecutive rows of the sorted matrix, which makes the binary searches on F and/or q-grams of the matrix very efficient. For k -mismatch problem, we record and keep only those groups for which the number of errors is less than or equal to k . Solutions for $k=0$ correspond to exact matches.

Tests were performed on different pattern lengths using 133 selected files from the Canterbury, Calgary, and TREC corpus. The results on k -mismatch pattern matching show that the running time and storage are superior to the fast suffix tree approach. The doctoral dissertation by Nan Zhang gives further details of our work (see publications).

3. A number of algorithms have recently been developed to search files compressed with the Burrows-Wheeler Transform (BWT) without the need for full decompression first. This allows the storage requirement of data to be reduced through the exceptionally good compression offered by BWT, while still allowing fast access to the information for searching. We provide a detailed description of five of these algorithms: Compressed-Domain Boyer-Moore, Binary Search, Suffix Arrays, q-grams and the FM-index, and also present results from a set of extensive experiments that were performed to evaluate and compare the algorithms. Furthermore, we introduce a technique to improve the search times of Binary Search, Suffix Arrays and q-grams by around 20%, as well as reduce the memory requirement of the latter two by 40% and 31%, respectively. Our results indicate that, while the compressed files of the FM-index are larger than those of the other approaches, it is able to perform searches with considerably less memory. Additionally, when only counting the occurrences of a pattern, or when locating the positions of a small number of matches, it is the fastest algorithm. For larger searches, q-grams provide the fastest results.

4. We first implement Amir's algorithm compressed domain pattern matching for LZ compressed files and make it practically useful by incorporating all the basic functionalities that a realistic pattern matching algorithm should possess viz. multiple occurrence of a pattern or multiple patterns matching in the compressed domain. We also report a faster implementation for so-called 'simple patterns'. Extensive experiments have been conducted to test the search performance and to compare with the BWT-based compressed pattern matching algorithms. The results showed that our method is competitive among the best compressed pattern matching algorithms. The LZW is a universal compression algorithm and our method requires no modification on the compression algorithm. We also worked on multiple-pattern matching in LZW compressed files using Aho-Corasick algorithm. The algorithm takes $O(mt+n+r)$ time with $O(mt)$ extra space, where n is the size of the compressed file, m is the size of the pattern length, t is the size of the LZW trie and r is the number of occurrences of the patterns. Extensive experiments have been conducted to test the performance of our algorithms. The results showed that our multiple-pattern matching algorithm is practically the fastest among all approaches when the number of patterns is not very large. Therefore, our algorithm is preferable for general

string matching applications. For further details see the doctoral dissertation by Tao Tao (see publications)

5. For image compression, we have studied algorithms for adaptive scanning-path for BWT-based lossless image compression. The methods use image statistics to predict the activity in the image. Based on this, and the nature of transformed output from the BWT, the algorithms determine the scanning path to use for the given part of the image. This provides adaptability in the scanning path without the time consuming problem of explicit edge detection or image segmentation. Details of this study can be found in the thesis by Nandakishore Jalumuri (see publications).

We also studied a new method to encode the prediction errors in lossless image compression. Using the BWT as a black box for image compression may not always produce a significant improvement in the compression results. Inspired by the BWT approach to data compression, we developed a new method for the encoding stage in prediction-based image compression systems. The encoding is based on the concept of alpha-paths, which provides a different view of the columns in the BWT rotation matrix. Initial results on natural images show that this method can significantly outperform most standard context-based image compression schemes. See the thesis by Rahul Parthe for details (see publications).

Training and Development:

Several Ph.D. and Masters students have participated and contributed in this research project, but not all of them received direct support from the grant. Individual Co-PIs meet with graduate students at their respective universities on a regular basis to discuss research problems. The students acquire the necessary skills to search literature and carry on an in-depth study and research in a field. The students are also asked to make presentations on their work. This gives the students experience of teaching graduate level courses and seminars. The overall effect of these activities is to train graduate students with the current research on the forefront of technology. Each one of them acquired valuable experience in undertaking significant programming tasks.

Outreach Activities:

Nothing yet to report

Journal Publications

N. Zhang, A. Mukherjee, D. Adjeroh and T. Bell, "Approximate Pattern Match Using the Burrows-Wheeler Transform", Proceedings Data Compression Conference, p. 458, vol. , (2003). Published

A. Firth, T. Bell, A. Mukherjee and D. Adjeroh, "A Comparison of BWT Approaches to String Pattern Matching", Software Practice & Experiences, p. 1, vol. 35, (2005). Published

Tao Tao and Amar Mukherjee, "Pattern Matching in LZW Compressed Files", IEEE Transactions on Computers, p. 929, vol. 54, (2005). Published

Tao Tao and Amar Mukherjee, "Multiple Pattern Matching in LZW Compressed Files Using Aho-Corasick Algorithm", Proceedings Data Compression Conference, p. 482, vol. , (2005). Published

Tao Tao and Amar Mukherjee, "Multiple Pattern Matching for LZW Compressed Files", Proc. International Conference on Information Technology: Coding and Computing, p. 91, vol. , (2005). Published

Tao Tao and Amar Mukherjee, "LZW Based Compressed Pattern Matching", Proc. Data Compression Conference, p. 568, vol. , (2004). Published

Tao Tao, Amar Mukherjee, Ravi Vijaya Satya, "A Search Aware JPEG-LS Variation for Compressed Image Retrieval", Proc. IEEE International Symposium on Intelligent Multimedia, Video and Speech Processing, p. 169, vol. , (2004). Published

Weifen Sun, Nan Zhang and Amar Mukherjee, "A Dictionary Based Multi-Corpora Text Compression System", Proc. Data Compression Conference, p. 448, vol. , (2003). Published

Books or Other One-time Publications

Nan Zhang, "Transform Based and Search Aware Text Compression Schemes and Compressed Domain Text Retrieval", (2005). Thesis, Published

Bibliography: Ph. D. Dissertation, University of Central Florida, School of Computer Science

Tao Tao, "Compressed Pattern Matching for Text and Images", (2005). Book, Published

Bibliography: Ph. D. Dissertation, University of Central Florida, School of Computer Science

Nandakishore Jalmuri, " Study of Scanning Paths for BWT-based Image Compression", (2004). Thesis, Published

Bibliography: MS Thesis, West Virginia University

Rahul Parthe, "Adaptive Edge-Based Prediction for Lossless Image Compression", (2005). Thesis, Published

Bibliography: MS Thesis, West Virginia University

Web/Internet Site

URL(s):

<http://vlsi.cs.ucf.edu/>

Description:

This site presents a complete description of our research and activities conducted under all the NSF sponsored research grants on data compression and compressed domain pattern matching. It posts all our publications electronically, it makes available all our annual and final reports submitted to NSF and people involved in the projects. If you follow the old website link (http://vlsi.cs.ucf.edu/old_root/index.html), it leads to M5 Online Compression Utility site where all our compression software and other compression software downloaded from different sites are made available online.

Other Specific Products

Contributions

Contributions within Discipline:

With the huge amounts of data often involved, efficiency considerations (in terms of both space and time) make it important to consider ways to keep the data in the compressed form for as much as possible, even when it is being searched. Our objectives in this proposal is to develop techniques for compressed domain pattern matching, i.e. to search for the required information directly on the compressed data, with minimal (or no) decompression. We proposed a class of new compressed domain pattern matching algorithms that exploits the sorted contexts of the Burrows-Wheeler transform. Our proposed methods are applicable to both text and images compressed based on the BWT.

Contributions to Other Disciplines:

Nothing significant to report.

Contributions to Human Resource Development:

Nothing significant to report

Contributions to Resources for Research and Education:

At the University of Central Florida, we have taught a graduate level course entitled 'CAP5937: Multimedia Compression on the Internet'. This has a new URL location: <http://www.cs.ucf.edu/courses/cap5015/>. This particular topic has grown directly out of the research that we have been conducting for the last four years on data compression. Lecture topics have included both text and image compression, including topics from the research on the current NSF grant. The course has been revised and is offered again Fall 2005. We offered this course four times so far.

At the West Virginia University, two graduate courses that relate to the project have been ongoing. EE558 Multimedia Systems have sections that discuss applications of compression to images and general multimedia data. EE568 Information Theory have sections that treat the fundamental basis and limitations of data compression. In the current report period, both courses (CS558, Fall 2003; EE568 Spring2004), have involved projects on lossless image compression, which are very relevant to the project. EE568 also involved projects and written reports on general data compression.

Contributions Beyond Science and Engineering:

Text searching is an important problem in diverse areas of human endeavor. With the emergence of the Internet, and the pervasive nature of email communication, we are just starting to appreciate the importance of fast text searching for both exact and inexact. With time, again thanks to the Internet and other improvements in communications and storage technology, images will become much more prevalent as they are today. And thus, people will want to search on images with the same ease that they use to search text data. Thus, the results from the proposed work will have impact far beyond the realms of computer science, or engineering, but in different aspects of our day to day activities as a society.

Categories for which nothing is reported:

Any Product