

Available online at <http://www.mecspress.net/ijwmt>

Big Data Compression in Mobile and Pervasive Computing

PankajDeep Kaur, Sandeep Kaur, Amneet Kaur

GNDU Regional Campus, Jalandhar

Abstract

The usability of Mobile devices and portable computers has been increased very rapidly. The main focus of modern research is to handle the big data in Mobile and pervasive computing. This paper gives an overview of Mobile Data Challenge which was a smart phone based research done by Nokia through Lausanne Data Collection Campaign. The Mobile Data Challenge was introduced when the amount of mobile data was seen to be increased very much in a short period of time. The rise of big data demands that this data can be accessed from everywhere and anytime. For handling such a huge amount of data e.g. SMS, images, videos etc, this data must be compressed for easy transmission over the network. It also helps in reducing the storage requirements. There are various techniques for the compression of data that are discussed in this paper. SMAZ and ShortBWT techniques are used to compress SMS. JPEG and Anamorphic Stretch Transform are used to compress the images.

Index Terms: Mobile Data Challenge; ShortBWT; SMAZ; Compression; Anamorphic Stretch Transform.

© 2016 Published by MECS Publisher. Selection and/or peer review under responsibility of the Research Association of Modern Education and Computer Science

1. Introduction

Now a day, there is a flood of data everywhere around. This data is in the of form videos, audios, images, emails, text messages, coming from the mobile devices, microphones, sensors, camera, social media etc, such a data is called Big Data. In 2009, the total data estimated to be 1ZB per year. In 2020, it is estimated to be 35ZB per year [1].

Big Data depends on the capabilities of the company. e.g. in Figure 1, 100TB of data is big data for company1 but not so for company2, because company 2 has capabilities to handle 100 TB of data so it is not Big Data for company2.

1.1 Mobile and Pervasive Computing

* Corresponding author.

E-mail address: pankajdeepkaur@gmail.com, kaurandeep116@gmail.com, arora.amneet@gmail.com

The idea behind pervasive computing is to develop a technology that is:

- Present everywhere and can be used anytime by the user (Omnipresent).
- Small to integrate to daily life easily e.g. mobile, watch, microwave.
- Working in background.

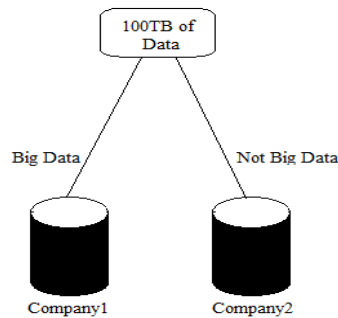


Fig.1. Big Data [1]

It was identified by Mark Weiser. He also named it as Ubiquitous computing (ubicom) [2].

Mobile Computing: It is an integration of portable computers as well as wireless network [15]. In this the connection is temporary and with time it is disconnected. Mobile computing is very active and evolving field of Research. 5 billion camera phones are there worldwide; Most of them have location awareness (GPS). 22% phones are smart phones. In 2013, numbers of smart phones were greater than number of PC's. These are big players in creating large amount of data [2]. Due to use of such number of cell phones, mobile networking is becoming huge. It introduced a project called Mobile Data Challenge (MDC).

2. Related Work

Juha K. Laurila, Daniel Gatica-Perez, Imad Aad, Jan Blom, Olivier Borne, Trinh-Minh-Tri Do, Olivier Dousse, Julien Eberle, Markus Miettinen [6] in 2012 presented an overview of LDCC (Lausanne Data Collection Campaign), this campaign raises MDC (Mobile Data Challenge) due to increase of data in a region. When the data was collected from the smartphones of around 200 users (participants of campaign) and that data was used for the research purpose then the MDC was introduced due to increase in the mobile data in that region. Protection of the participants personal data and privacy is their main concern, that why the collected data is made available to the researchers in a limited manner.

Pankaj Bhaskar and Sheikh I Ahamed [7] in 2007 explained privacy as one of the major issue in pervasive computing and proposed some models to deal with privacy challenges. The main reason why this issue raise is because of limited understanding of the user about the technology. The privacy issue arise due to large increase of data due to increase in number of pervasive computing devices and mobile devices.

Bo Li and Prof. Raj Jain [8] in 2013 proposed the cloud and user's effective interaction. Mobile networking is becoming larger due to large increase in number of cell phones or smart phones. They introduced benchmarks as well as progress in mobile networking.

Minos Garofalakis, Kurt P. Brown, Michael J. Franklin, Joseph M. Hellerstein, Daisy Zhe Wang, Eirinaios Michelakis, Liviu Tancau, Eugene Wu, Shawn R. Jeffery, Ryan Aipperspach [7] in 2006 explained data furnace project that is built at UC-Berkeley and Intel Research for data management in pervasive computing. It is the central repository of metadata (data about data) as well as application data.

Mohammad H. Asghari, and Bahram Jalali [10] in 2014 presented the JPEG and AST compression and show the improved compression when JPEG is followed by AST compression by computing the structural similarity between the images. The image has more resolution in this case.

Paul Gardner-Stephen, Andrew Bettison, Romana Challans, Jennifer Hampton, Jeremy Lakeman, Corey Wallis [11] in 2013 presented that compression of Short SMS by various techniques and explains how short messages are difficult to compress than longer messages.

3. Mobile Data Challenge

MDC was a smart phone based research done by Nokia Company through Lausanne Data Collection Campaign (LDCC). It was started in January, 2009 by Nokia Research Centre, Idiap and EPFL. In this Campaign, the data from smart phone of 200 volunteer people in the region of Lake Geneva was collected over one year of time. The data collection was started in October 2009 by using Nokia N95 phones. The data collection is made invisible to participants by client server architecture and the data was recorded on a database [6].

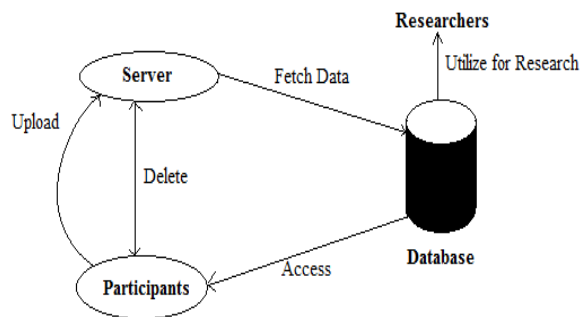


Fig.2. LDCC Data Flow [6]

The data was first uploaded to the server via Wireless Local Area Network (WLAN). The data received by the server is recorded in the database. This data was accessed by the participants of the campaign. The server was Simple Context server that was developed in Palo Alto by Nokia. The data from database was accessed or utilized by the researchers for research purpose on it. The type of data recorded was the data related to phone calls, SMS logs, Bluetooth, GPS, user downloaded applications, audio, videos, photos etc. LDCC was done in 2009-2011 and the challenge arises in 2011-2012 when amount of data is seen to be increased a lot, thus named this Mobile Data Challenge. In this, every researcher was committed for the privacy of user's data to only treat it for research purpose. They were committed to respect the user's privacy. The collected data or big data are made available to the researchers in a limited manner due to privacy concern [6].

4. Pervasive Computing

A wide number of sensors are used in pervasive computing, from richly semantic and high bandwidth sensors to Boolean sensors i.e. from video cameras to door-ajar switches. In industries, organisations run analytics for their business processes on sensor data. The nature of pervasive computing is "real-time", they need to store large volumes of data due to increase in number of pervasive applications e.g. sensors. Idea brought by pervasive computing is Invisibility i.e. technology should be made so common and it should continue working to the user's expectation and present him with little or no surprise such that he will not notice that there is any technology existing behind e.g. sensors embedded inside floor [8]. Large volume of data

i.e. big data is needed in various pervasive computing projects as aware home project, guide projects, My life Bits project etc [2].

- Aware home project: focussed on creating a smart environment at home in which everything that a person, who lives in, does is recorded. Small processors are embedded inside the whole house. Information recorded can be later used to investigate how people behave at home, how they live. It later become the project in which assistance of older people is done in home, as with time the memory fades away a bit.
- Guide projects: device that helps us to navigate a city, using wireless network. This device will tell where we are at present time and shows what is around us and shows routes to destination e.g. Google maps.
- My Life Bits Project: In this, everything we done in our day can be recorded i.e. pieces of your life can be recorded and saved inside computer [2].
- Google Glasses: It is development by Google. It came from the idea of pervasive computing i.e. it include wearable computing and augment reality. Wearable computing means an embedded chip and augment reality means one's idea with other people idea gives better result. These are the eye glasses that show the information from internet onto the lenses that uses a 4G cell connection. It communicates with the mobile phones through the Wi-Fi and then displays the contents on video screen and also responds to the voice commands of the user. It consists of video display, camera, microphones, buttons, speakers etc [13].

There is one project data furnace project that is built at UC-Berkeley and Intel Research for data management in pervasive computing. It is the central repository of metadata (data about data) as well as application data. It has mainly three layers: *Hardware Layer* that manages physical resources used for communication, storage and processing. It contains CPU, Storage, Network etc. *Metadata Layer* contains Database Schema, Event Definitions, Application Programming Interface and devices. *Service Layer* contains Data Archiving and Streaming, Event Processing, Data Model Learning etc. It is best suitable for application like smart home[9].

5. Compression

Compression is must now a days, with the increase of big data. There are huge amount of pervasive computing applications that produces large amount of data. Compression helps in reducing the storage requirements. By compression, data takes less space, so more data can be stored in same space and also due to small size it is easier to transmit it over the network. Following are the some big data compression techniques:

5.1 SMAZ

It is a lossless compression technique used for the compression of small strings, also for SMS (Short Message Service). It is a compression library that compresses every type of text data. In general short messages are difficult to compress than longer messages because a longer message has sufficient redundancy that helps to reduce its size through a compression scheme easily. Therefore for short message compression, the logical solution is to pre-compute a kind of dictionary.

SMAZ is an open source compression library. The complete implementation of SMAZ is less than 200 LOC. It claims to reduce the size of message of up to 50% - 60% of original size. For Example: "The" is compressed to a single byte. SMAZ is also used for the compression of URL's as well as short HTML files. There are two functions in the library: Compression function and Decompression function.

```
int smaz_compress(char *in, int inlen, char *out, int outlen);
```

It is the compression function where 'in' is the input buffer and 'inlen' is the length of the buffer, 'out' is the output buffer of length 'outlen' where the compressed data is put. The decompress function is:

```
int smaz_decompress(char *in, int inlen, char *out, int outlen);
```

This function decompresses the buffer 'in' having length 'inlen' and then put decompressed data in the buffer 'out' of maximum length 'outlen' bytes. If output buffer 'out' is too short that it can't hold decompressed string, then outlen+1 is returned. Else, the length of string compressed is returned. Table 1 gives the performance of SMAZ on various types of inputs. Percentage is compressed size of original image. Lower the percentage better is the compression. E.g. in twitter messages, the compression is worst with percentage 81.4% [11]

Table 1. Performance of SMAZ compression for various inputs [11]

Corpus	Type	Messages	Avg. Length	SMAZ
Pie Floater	English	28	171	60.6%
SMAZ README file	English	71	44	64.1%
British English SMS	SMS	450	80	64.8%
SMS SPAM Collection v.1	SMS	5,547	82	71.6%
Private Twitter Corpus	Twitter	348,994	54	81.4%

5.2 ShortBWT

This technique is also used for the compression of Text. It does not use any precomputed dictionary like SMAZ. It is based on properties of the Burrows Wheeler Transform (BWT). The shortBWT compresses 200 characters strings to 55% of original size [11].

There are four stages of BWT namely- BWT (Burrows Wheeler Transform), GST (Global structure transformation), RLE (Run Length Encoding), EC (Entropy coding). Output of one stage is the input of the next stage and processes from left to right.

The first stage is BWT- It includes block sorting algorithm in which text is divided into equal sized blocks with each block processed separately from one another. The blocks are sorted using block sorting algorithm. These blocks generated uses same symbols but in different order. This will group the same or repeating symbols [18].

The second stage is GST- It includes MTF (Move to First) algorithm. MTF rearranges the pattern such that more frequently used symbols tend to be near the front of the string and less frequently used symbols towards the end [17].

The next stage is RLE- It shrinks the number of symbols by Run Length Encoding. The last stage EC produces a bit output with the help of arithmetic coding scheme.

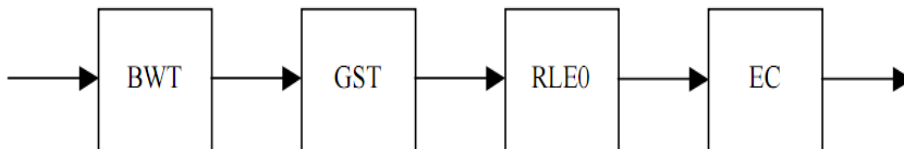


Fig.3. BWT compression steps [18]

5.3 JPEG

Increase in the number of pixels of image sensors is responsible for big data. JPEG is the technique used to compress image.

The steps involved in image compression are:

- Divide image into 8x8 pixels sized blocks.
- Apply DCT: Discrete Cosine Transform (DCT) is applied to each block and the value of each pixel is converted to frequency domain. The lower frequencies appear at Top-left side of block and higher frequencies appear at bottom-right side. Thus, it separates the information that is more noticeable from the information that is less noticeable. IDCT (Inverse Discrete Cosine Transform) is applied to reform the original image.
- Quantization: used for discarding insignificant information. It divides the each 'DCT coefficient' by 'quantization coefficient' and returns an integer value by discarding the value after decimal point. Each DCT values have their own quantization coefficient. Quantization coefficient is chosen in a way such that there may be not any noticeable change even after the value following the decimal is discarded.
- Encoding: After the quantization step, most of the DCT coefficient have redundant amount of data. Huffman coding will remove this redundant amount of data. The Huffman coding is a type of Variable-length coding, in which the symbols that occur more frequently are encoded by using fewer bits and the symbols that occur less frequently are encoded by more bits and all this information is stored in a look-up table [15][16]. The image obtained after these steps is a compressed image and it is converted to original image by applying decoding, de-quantization and Inverse DCT.

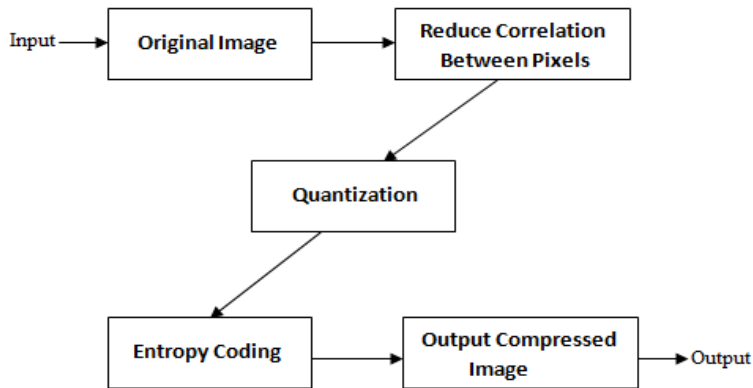


Fig.4. JPEG compression steps [15]

5.4 AST

To enhance the performance of JPEG, AST is introduced by UCLA researchers for the compression of digitized data. AST is Anamorphic Stretch Transform. AST increases coherency of waveform so when it is sampled, the sampling rate become less than the Nyquist rate and thus it reduces the size. The data is compressed by stretching and wrapping data by mathematical function. In this the quality of data is maintained as compared to other compression techniques. It is used today as a solution to the 'Big Data' problem. AST works in analog as well as digital domain. In digital records as medical data, this transformation reshapes the signal in which the 'coarse' features are stretched less than 'sharp' features

$$B'[n,m]=|\sum e^{j.\phi[k1,k2]} .B[n-k1,n-k2]|$$

Where $B'[n,m]$ - original image brightness, n and m represent 2D discrete spatial variables, $\phi[n,m]$ -nonlinear phase profile and $|\cdot|$ - nonlinear absolute. Here, $k1$ and $k2$ lies between $-\infty$ to ∞ . If AST is done after JPEG compression, the resolution is higher than the resolution in JPEG compression alone even if the compression factor is same. To compare the performance of Image compression Structural similarity (SSIM) is computed. SSIM for JPEG compression is 87.2% but with JPEG along with AST is 95.3%. Therefore performance is improved if JPEG is followed by AST [10].

In many compression techniques there is more loss of data when the image is compressed. Quality of image is lost when compressed. With this technique, the data quality is maintained. It wraps image such that stretching of small features is done rather than large features. Thus small features are assigned with more samples (bits) and less samples (bits) to large features in which data is redundant. So the result is a smaller file size having high resolution so that the object or image can be seen clearly [17].

6. Conclusions

In this paper, we conclude that growing number of mobile devices, microphones, sensors etc are the main data generating points. Mobile Data Challenge (MDC) arises due to increase in mobile data through Lausanne Data Collection Campaign (LDCC) by Nokia. LDCC was done in 2009-2011 and the challenge arises in 2011-2012 when amount of data is seen to be increased a lot. Huge amount of data need extensive effort for collecting it, so it need to be compressed. Compression technique SMAZ and ShortBWT are used to compress SMS. SMAZ is a compression library that is used to compress very small strings. In an average, SMAZ compresses data by 40-55% and ShortBWT compresses 200 characters strings to 55% of original size. The image compression techniques are JPEG and Anamorphic Stretch Transform (AST). If AST is done after JPEG compression, the resolution is higher than the resolution in JPEG compression alone even if the compression factor is same. The performance is improved if JPEG is followed by AST compression.

Acknowledgement

We are very grateful to our advisor Pankaj Deep Kaur for her support, patience, motivation and immense knowledge. We could not have imagined of having a better advisor than her. Her guidance helped us in all the time of writing this survey paper.

References

- [1] www.youtube.com/watch?v=PlaJsseTgk4
- [2] www.research.nokia.com/mdc.
- [3] www.en.wikipedia.org/wiki/Video_compression_picture_type
- [4] www.cse.iitb.ac.in/synerg/lib/exe/fetch.php?media=public:proj:ganesh_video_adaptation:introduction-to-video-compression.pdf
- [5] Juha K. Laurila, Daniel Gatica-Perez, Imad Aad, Jan Blom, Olivier Bornet, Trinh-Minh-Tri Do, Olivier Dousse, Julien Eberle, Markus Miettinen “ The Mobile Data Challenge: Big Data for Mobile Computing Research” in 2012.
- [6] Pankaj Bhaskar and Sheikh I Ahamed “Privacy in Pervasive Computing and Open Issues” in Proceedings of the International Conference on Availability, Reliability and Security (AREs), IEEE CS Press, Vienna, Austria, April 2007.
- [7] Bo Li and Prof. Raj Jain ”Survey of Recent Research Progress and issues in Big Data” in 2013.

- [8] Minos Garofalakis, Kurt P. Brown, Michael J. Franklin, Joseph M. Hellerstein, Daisy Zhe Wang, Eirinaios Michelakis, Liviu Tancau, Eugene Wu, Shawn R. Jeffery, Ryan Aipperspach, “Probabilistic Data Management for Pervasive Computing: The Data Furnace Project” in 2006.
- [9] Mohammad H. Asghari, and Bahram Jalali” Discrete Anamorphic Transform for Image Compression” in May 27-31, 2014
- [10] Paul Gardner-Stephen, Andrew Bettison, Romana Challans, Jennifer Hampton, Jeremy Lakeman, Corey Wallis “ Improving Compression of Short Messages” in May, 2013
- [11] www.youtube.com/watch?v=3twBv2v4Ip0
- [12] www.lesliefisher.com/handouts/glass_fisher.pdf
- [13] www.cse.wustl.edu/~jain/cis788-95/mobile_comp/
- [14] A.M.Raid, W.M.Khedr, M. A. El-dosuky and Wesam Ahmed “Jpeg Image Compression Using Discrete Cosine Transform - A Survey” in April 2014
- [15] www.math.tau.ac.il/~turkel/notes/JPEG.pdf
- [16] www.theaggie.org/2014/01/23/warping-can-compress-big-data/
- [17] Radu R. ADESCU “Transform Methods Used in Lossless Compression of Text Files” in 2009.
- [18] Juergen ABEL “Improvements to the Burrows-Wheeler Compression Algorithm: After BWT Stages”

Authors’ Profiles



Dr. Pankaj Deep Kaur is working as an Assistant Professor in the Department of Computer Science and Engineering, Guru Nanak Dev University, RC, Jalandhar, India. She received her Bachelor’s Degree in Computer Applications (2000) and Master’s Degree in Information Technology (2003) from Guru Nanak Dev University, Amritsar, India. She completed her Ph.D. in Resource Scheduling in Cloud Computing from Thapar University, Patiala (2014) and has over ten years of teaching and research experience. She has been a university position holder in her graduation studies and received Gold Medal for her excellent performance in her Post Graduation studies. She is a recipient of Junior Research Fellowship from Ministry of human Resource and Development, Govt. of India. Her research interests include Cloud Computing and Big data.



Sandeep Kaur is currently pursuing her post graduation from Guru Nanak Dev University Regional Campus Jalandhar. She received her Bachelor’s Degree in Information Technology (2014) from Baba Banda Singh Bahadur Engineering College, Fatehgarh Sahib.



Amneet Kaur is currently pursuing her post graduation from Guru Nanak Dev University Regional Campus Jalandhar. She received her Bachelor’s Degree in Computer Science (2014) from Guru Nanak Dev Engineering College, Ludhiana