Journal of Information Engineering and Applications
ISSN 2224-5782 (print) ISSN 2225-0506 (online)
Vol.8, No.1, 2018
www.iiste.org

Big Data Analysis: A New Scheme for the Information Retrieving Based on the Content of Multimedia Documents

Abderrahmane EZ-ZAHOUT
Higher National School of IT / ENSIAS College of Engineering, Rabat, Mohammed V University, Morocco

Jawad OUBAHA
Higher National School of IT / ENSIAS College of Engineering, Rabat, Mohammed V University, Morocco

Abstract
Big Data analysis is one of the hot topics now days for knowledge discovery in databases process. It's considered as significant field of knowledge management. Roughly, the ¾ of organizations have been adopted some form of analytics today. The most posed question in big data analysis is how to manage and operate in it? In this study, we explain the concept of the proposed information system architecture for retrieving information. This system scheme operates basing on the content of the document. Digitized visual media: images and videos captured from real time video surveillance system require high storage capacity. This work describes the steps of indexation and content modeling for retrieving and managing information in multimedia documents databases.
Keywords: big data analysis, multimedia documents, indexing, modeling, classification, content representation.

1. Introduction

Recently, big data analysis is an emerging data science paradigm of business analytics and of multi dimensional information mining. The big quantity of information collected from various scientific explorations and transactions often tools to assist efficient data analysis, management and retrieving rapidly and efficiently interest information. The big data problem has become a serious issue in the society. The big data is characterized by the Volume, Velocity, Variety and Veracity [1].

The big data challenge is the retrieving information topics used especially for correct making decision. For example in a real time video surveillance system as designed and developed in [2], a big multimedia data waits for treatment and an operation of intervention doesn't be false. The system should be able to correctly analyzing and retrieving the abnormal profile in real time.

All video streams should be saved and scanned by the detection, tracking, profile analysis and re-identification of the non ordinary person. During running of the process of features extraction, large volume, variety forms and very changing data sets are collected from connected cameras and machine learning techniques can be used to retrieve and analyze the relevant information.

Another example of web information analytics in which retrieving related and significant information for some themes like "terrorism" is important. In these systems several web sites explore biggest quantity of information and a large range of content is present. Increasingly, companies should be able to know limitations of their sites and to mine their data. Doing this requires efficient tools information systems to perform analytics and to manage big data for not leasing in transactions.

We hope that in 2030, no house, no company, no street, and no car and in each places there are cameras installed. And information should be stored and managed to make decision in a time.

Sadly, in front of various textual data, static images and videos can't offer a readable semantics information.

A multimedia information system should exist to extract relevant information and and deduct a semantic information. To achieve this indexation process and searching for visual media process are required.

1.1 Document modeling

This consists on a formal representation that facilitates the structuring data in an information system. In this paper we are interested into visual multimedia information like image and video. For images many mathematical tools are used for modeling images as semi-structured objects: FFT, wavelets and others. In other hand the modeling of a video document consists of organizing clearly of its content. Contrary to classical documents, we should use the visual and temporal and then in some streams the sound components. Generally, an image document is represented as described as:
For I class of images:

$$I = C^{\mathbb{N}}$$

Where C is a set of polymorph of characteristics [3] In modeling images, three level of data can be used:

1. Low level data : color and texture,

Intermediate level data: a segmentation of images is used to retrieve more significant information of each bloc of image.

2. High level data: Indexation is used for providing semantic descriptions of the scene and objects in the

real world.

For the video searching systems by content are based on three steps: structuring, annotate and organizing in order to utilize annotate documents [7]. For example, a frame of a video sequence is presented by a set of objects and relations among them and each one of the objects has many properties including media features, visual and semantic features.

1.2 Multimedia documents database modeling

Generally speaking, two types of multimedia documents exist, the simple multimedia documents and the complex multimedia documents. The first ones comprise an instance of one type of media while the second comprise several instances of one or many types of media.

1.3 Searching information by the content

Clearly, searching information in normalized DB is well controlled and very easy in implementation. But, in multimedia DB involves a lot of problems.

The most known are the definition of queries and high level data. The first is as important then the second. Many structures of searching are defined in the literature like searching by formal queries, feedback searching and searching by navigation.

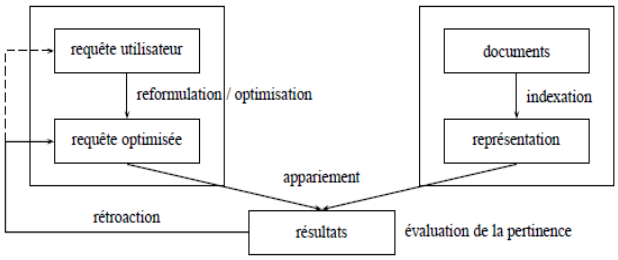

Figure 1: Research information general scheme [4].

This figure represents the video meta-modeling system. This system is based on video structuration and on its annotation.

2. Big Data- Visual Media Modelisation

The big data model is an used to manage the data stored in physical devices. Actually, we have large volumes of data with different formats stored in different ways.

The big data modeling provides a visual way to analyze data resources, and creates fundamental data architectures. This can help have more applications to optimize data reuse and reduce computing costs [5].

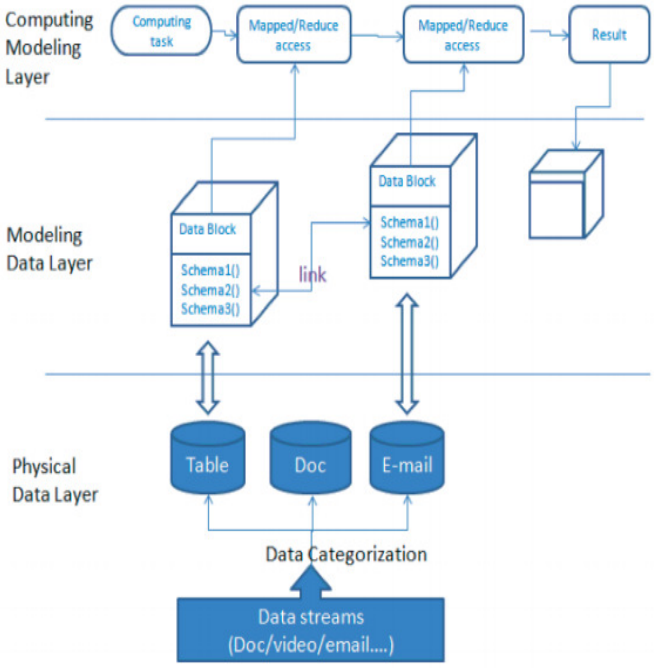

Figure 2: General architecture of big data [5].

This figure explain that the effect to construct a big data model, we must first create data blocks based on data

type, read-write requirement, data storage and relationship.

2.1 Multimedia documents

As definition, multimedia documents are natural extensions of a conventional textual document in the multimedia area. There are defined as a digital document composed of multiple types of media elements (text, image, video, sound etc.).

In term of big data analysis, multimedia documents are the very important issue that costs more and contains very significant information.

3. Overview of the Proposed Information System Architecture

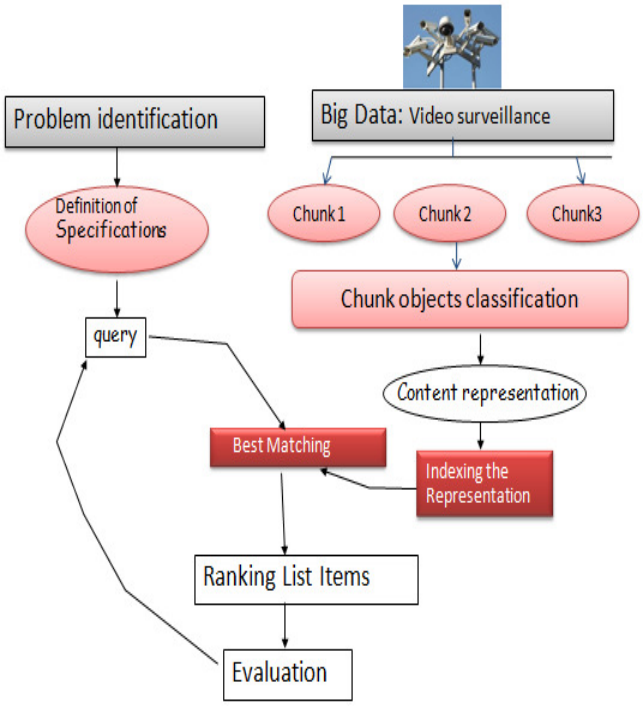

Figure 3: The proposed system overview

In this architecture, suppose that the multimedia documents are received from a network of connected cameras in a video surveillance system. The big data stored and created over the real time working, is cutted to different documents chunks (images, sound, text, background) and then each class of document type are classified to identify types of objects of the big data. EM [6] is the flagship algorithm is proposed for this classification. Then an indexation is made for the contents in order to search the best matching between indexes.

An evaluation of the effectiveness of this system is proposed. This evaluation is based on two most commonly known metrics:

The effectiveness of retrieving documents is made by:

$R_{effecf} = \frac{No.of\ the\ relevant\ documents\ retrieved}{No.of\ the\ documents\ in\ the\ DB}$ The metric of precision is calculated as:

$P = \frac{No.of\ the\ relevant\ documents\ retrieved}{No.of\ the\ documents\ retrieved}$ In Multimedia documents management, the problem of measurement in information retrieval has an important aspect. The basic variables underlying any measure of retrieval effectiveness are usually Expected Precision (P) and Expected Recall (Reffect).

All calculations of Expected Recall and Expected Precision are based on the assumption that the set of relevant documents for a query is the same, no matter who the user is. Then a similarity measurement between indexes of the content representation data base by:

$$\text{Similarity}(I_q, I_j) = S_j, \vec{d}_j) = \frac{\vec{q} \bullet \vec{d}_j}{|\vec{q}| \times |\vec{d}_j|} = \frac{\sum_{k=1}^{t} w_{qk} w_{jk}}{\sqrt{\sum_{k=1}^{t} w_{qk}^2 \sum_{k=1}^{t} w_{jk}^2}}$$

With:

$\vec{q}$ is the vector of a given query and is the vector of the j[th] index of the document.

wi are weighted of indexes and $\vec{d}_j$ represents the importance of index *i* in the multimedia document. The main problem of Information research and retrieving in the multimedia document is the extraction of the specifications and the representation of the documents.

This representation should be amenable to automatic processing. This representation concerns the extraction of descriptors, indexing of each chunk and extraction of special forms. And finally, queering in a language understood by the information system.

This scheme offers a new process succession for information retrieving. This diagram is based on cutting a big document to manageable chunks. This technique allows short memory access time to read information from and in. and all the sub-process are executed rapidly.

4. Conclusion

This research provides a new scheme for retrieving information in very big information system. And especially address to manage big data in multimedia forms. The proposed system can be paralleling used with a intelligent network on connected camera in a video surveillance system. It describes the steps of indexation and content modeling for retrieving and managing information in multimedia documents databases.

5. References

1. Wolff, J. G. (2014). Big data and the SP theory of intelligence, IEEE Access, 2, 301-315.
2. Ezzahout A. and Oulad Haj Thami R., Conception and Developpement of video surveillance system for detecting, tracking and profile analysis of a person, in the 3d. International Symposium ISKO-Maghreb'2013 "Concepts and Tools for Knowledge Management", November 8th - 9th, Marakech – Morocco.
3. E. Bertino, B. C. Ooi, R. Sacks-Davis, K. L. Tan, J. ZobelL, B. Shidlovski et B. Catania. Indexing Techniques for Advanced Database Systems. Kluwer Academic Publishers, Boston, 1997. 250 pages.
4. I. Mbaye, R. Oulad Haj Thami, J. M. Martínez: A Model for Indexing Videos and Still Images from the Moroccan Cultural Heritage. MMSP 2005: 1-4.
5. Jinbao Z., Data Modeling for Big Data, International Journal of Computer Science and Applications, Vol. 12, No. 1, pp. 1 – 15.
6. Andrew N., The EM algorithm, Part IX, CS229 Lecture notes.
7. Velmurugan L., P.Sasikumar, Alema Gebre and Tilahun.A, Big Data Analysis and Its Applications for Knowledge Management, International Journal of Computer Science and Information Technology, 2015, 1(1),1-5.

Authors Profile

Dr. Abderrahmane Ez-zahout has received the Phd degrees in computer sciences, from Higher School of IT/ ENSIAS college of Engineering – Mohammed V University of Rabat, Morocco. Respectively, degree in advanced studies DESA. He developed research activities covering video processing and Big Data Analysis related fields.

Dr. Jawad Obaha is a Doctor in computer sciences from Higher School of IT/ ENSIAS college of Engineering – Mohammed V University of Rabat, Morocco. Respectively, degree in advanced studies DESA. He developed research activities covering Computer Networks and Big Data in information systems related fields.