Computers in Biology and Medicine 66 (2015) 190–208

Contents lists available at ScienceDirect

Computers in Biology and Medicine

journal homepage: www.elsevier.com/locate/cbm

A similarity-based data warehousing environment for medical images

Jefferson William Teixeira [a], Luana Peixoto Annibal [c], Joaquim Cezar Felipe [b], Ricardo Rodrigues Ciferri [c], Cristina Dutra de Aguiar Ciferri [a,*]

[a] Department of Computer Science, University of São Paulo at São Carlos, 13.560-970 São Carlos, SP, Brazil
[b] Department of Computing and Mathematics, University of São Paulo at Ribeirão Preto, 14040-901 Ribeirão Preto, SP, Brazil
[c] Department of Computer Science, Federal University of São Carlos, 13.565-905 São Carlos, SP, Brazil

ARTICLE INFO

Article history:
Received 13 March 2015
Accepted 25 August 2015

Keywords:
Data warehousing
Medical database
Medical image
ETL process
OLAP query processing
Similarity search
Indexing
Bitmap index

ABSTRACT

A core issue of the decision-making process in the medical field is to support the execution of analytical (OLAP) similarity queries over images in data warehousing environments. In this paper, we focus on this issue. We propose *imageDWE*, a non-conventional data warehousing environment that enables the storage of intrinsic features taken from medical images in a data warehouse and supports OLAP similarity queries over them. To comply with this goal, we introduce the concept of perceptual layer, which is an abstraction used to represent an image dataset according to a given feature descriptor in order to enable similarity search. Based on this concept, we propose the *imageDW*, an extended data warehouse with dimension tables specifically designed to support one or more perceptual layers. We also detail how to build an *imageDW* and how to load image data into it. Furthermore, we show how to process OLAP similarity queries composed of a conventional predicate and a similarity search predicate that encompasses the specification of one or more perceptual layers. Moreover, we introduce an index technique to improve the OLAP query processing over images. We carried out performance tests over a data warehouse environment that consolidated medical images from exams of several modalities. The results demonstrated the feasibility and efficiency of our proposed *imageDWE* to manage images and to process OLAP similarity queries. The results also demonstrated that the use of the proposed index technique guaranteed a great improvement in query processing.

1. Introduction

The decision-making process plays an important role in the medical field. Based on summarized and integrated data, specialists (e.g. hospital superintendents, board members, and medical staff leaders) are able to make more strategic and productive decisions concerning the clinical routine. For instance, the analysis of medical information can support efficient planning strategies against disease outbreaks (including the identification of contagious geographical areas or communities), the establishment of vaccination campaigns, the improvement in the vaccination delivery services, the development of educational campaigns, and the analysis of treatment effectiveness.

Data warehousing environments (DWEs) have emerged as a core component for the decision-making process, providing high quality informational data. Through the ETL (extract, transform, load) process, DWEs consolidate large amounts of data of interest from heterogeneous and distributed data sources into a specialized database, the data warehouse (DW). Besides being integrated, data in the DW are subject-oriented, multidimensional, historical and non-volatile [1]. Also, DWs are often implemented in relational databases through a star schema, which is composed of fact and dimension tables. Fact tables store numeric measures of interest while dimension tables contain attributes that contextualize these measures. Queries on such DWEs are called OLAP (on-line analytical processing), and they enable decision-making users to issue analytical queries against the DW without the need of accessing the original data sources [2].

Conventional DWs store only conventional data, such as data of numeric, alphanumeric, and date types. As a result, conventional DWEs only support the ETL process and the OLAP query processing based on conventional data. Conventional data types have a total order relation, and therefore can be sorted and searched using ordinary relational operators (i.e. $<, \leq, >, \geq$). However, decision-making users cannot use currently conventional DWEs to issue OLAP queries over complex multimedia data, such as images. This is related to the fact that, differently from conventional data, complex data do not have a total order relation, and therefore cannot be sorted and searched using ordinary relational operators.

* Corresponding author. Tel.: +55 16 3373 8172; fax: +55 16 3361 7906.
E-mail addresses: williamteixeira5@gmail.com (J.W. Teixeira), annibal.l.p@gmail.com (L.P. Annibal), jfelipe@ffclrp.usp.br (J.C. Felipe), ricardo@dc.ufscar.br (R.R. Ciferri), cdac@icmc.usp.br (C.D.d.A. Ciferri).

http://dx.doi.org/10.1016/j.compbiomed.2015.08.019
0010-4825/

In Example 1, we introduce a case study that illustrates different analyses that use conventional DWEs in the medical field. We also highlight an analysis that cannot be performed, as it requires the management of image data. This case study is used throughout the paper.

Example 1. Suppose a conventional DWE that integrates clinical data related to different modalities, such as breast, head, and knee related problems. These exams were collected over several years, and were carried on patients from different hospitals, belonging to different age groups. Based on the conventional data stored in the DW, specialists can perform the following OLAP analyses: simple trend analysis (e.g. "What is the incidence of breast cancer in 2011 in the southeast region of the US?"), comparative analysis (e.g. "What is the incidence of breast cancer in the last 3 years in the southeast region of the US?"), and multiple trend analysis (e.g. "What is the incidence of breast cancer in the last 3 years in the southeast region of the US, considering different age groups?").

However, there is a range of queries that cannot be issued against this conventional DWE. For instance, specialists cannot use this application to evaluate the prevalence of certain types of pathology through OLAP queries like: "How many images, similar to a specific breast cancer image, occurred in patients from the southeast region of the US, aged between 30 and 40 years old, in the last 3 years?". This kind of analysis involves comparisons over images, which are not supported by conventional DWEs.□

Managing only conventional data in DWEs has two main drawbacks. First, there is a crescent volume of medical images produced day-by-day in clinics and hospitals, which have to be stored separately from the DW. Second, it is not feasible to use these images in the OLAP query processing to perform similarity search in a tandem with conventional data. This impairs the decision-making process, as specialists cannot base their decisions on important information that can be obtained from images.

The drawbacks motivate the challenge of expanding the storage capacity of DWEs (and their respective ETL and query processing) to support images. Regarding the OLAP query processing, a core challenge is to reduce the "semantic gap" in queries based on similarity search. The semantic gap refers to the difference between the result produced by a computational query and the result expected by specialists [3,4]. In fact, different specialists may choose different aspects to determine the similarity among a set of images, according to the purpose of the ongoing medical task. That is, they may choose to analyze images according to different feature descriptors, which describe the intrinsic features taken from images (i.e. the images visual patterns) mostly regarding color, texture, and shape. For instance, a given specialist may choose to analyze the texture feature of images using the Haralick descriptors [5], while another specialist may choose to analyze the color feature of images using Colors Histograms [6]. Thus, the precision of similarity comparisons depends on the task and on the specialist's perception.

In this paper, we focus on these challenges. We propose *imageDWE*, a non-conventional DWE that enables the support of medical images in the data warehouse and the processing of OLAP similarity queries over these images. Using our proposal, specialists can perform a new range of interesting analyses, such as that described in Example 1.

We have designed *imageDWE* to provide major characteristics as follows:

- It defines how images should be stored in the DW. To this end, we extend the star schema design of conventional DWs to also encompass dimension tables specifically designed to store data related to images. Thus, our proposed *imageDW* is able to store conventional and image data together.
- It reduces the semantic gap in similarity queries. To comply with this goal, we introduce the concept of perceptual layer, which is an abstraction used to represent an image dataset according to a given feature descriptor (e.g. Color Histograms, Haralick) in order to enable similarity search. Each perceptual layer represents a particular specialist' perception. All perceptual layers of interest are stored in the *imageDW*.
- It extends the ETL process to manage images. That is, we empower the conventional ETL process to also generate data for the perceptual layers and to store them in the *imageDW* accordingly.
- It extends the OLAP query process to support similarity queries over images. To comply with this goal, we integrate the conventional OLAP and the similarity search processes.
- It introduces an index technique, which encompasses the specification of an index and the definition of different processing strategies. The technique improves even more the extended OLAP processing of similarity queries over images provided by *imageDWE*.

We have presented a preliminary version of this study in [7]. Here, we extend that work by allowing the storage of several perceptual layers in the DW, and by extending the conventional ETL process and OLAP query processing to support these perceptual layers. This allows specialists to perform more complex analyses based on different aspects of images. Furthermore, we introduce a novel index technique to support the extended OLAP query processing. Moreover, we describe new performance tests that highlight the advantages of our proposal.

This paper is organized as follows. Section 2 describes the background needed to understand our proposal, Section 3 introduces the proposed *imageDWE*, which is described in terms of the star schema of the *imageDW* and the extended ETL and OLAP query processing capabilities, Section 4 introduces the index technique, Section 5 provides a set of guidelines for expanding *imageDWE*, Section 6 describes the performance evaluation of *imageDWE*, Section 7 surveys related work, and Section 8 concludes the paper.

2. Background

We detail the main issues related to similarity search over images in Section 2.1. Also, our work is based on two well-known concepts available in the literature: the Omni technique and the star-join Bitmap index, which are described in Sections 2.2 and 2.3, respectively.

2.1. Similarity search

To be computationally analyzed, images should be pre-processed using through feature extractors, which are responsible for generating feature vectors that describe their intrinsic characteristics [8]. This process is detailed as follows. An image is represented as a two-dimensional $m \times n$ matrix of pixels, where m and n are the image dimensions and the pixel have integer values that depend on the image type. For instance, the pixel values can be 0 or 1 in binary images, vary between 0 and 255 in grayscale images represented by 8 bits, and have three values in the range of 0–255 each in RGB color images.

A feature descriptor is characterized by [9]: (i) a feature extractor algorithm, which tracks down the images, processes their pixel values, produces numeric representations of them, and stores these values in feature vectors; and (ii) a distance function, which produces a similarity measure that is used to determine the