IADIS International Conference

e-Society 2003

Lisbon, Portugal

3-6 June 2003

IMAGE DATA MINING – CONCEPTUAL OVERVIEW OF IMPLEMENTATION AND METHODOLOGY CHALLENGES

Tamara Dinev
Florida Atlantic University
Boca Raton, FL, USA
tdinev@fau.edu

Peter L. Stanchev
Kettering University
Flint, Michigan 48504, USA
pstanche@kettering.edu
www.kettering.edu/~pstanche

ABSTRACT

Data mining is a class of analytical techniques that examine a large amount of data to discover new and valuable information. Data mining applications have proved highly effective in addressing many important business problems. Intelligent access to rich archive of image data in different fields makes the image data mining applications of particular interest. This paper delivers a conceptual overview of implementation and methodology challenges in image data mining. Different applications are discussed as well as the challenges specific to the nature of image processing and data mining.

KEYWORDS

Multimedia data mining, Knowledge discovery

1. INTRODUCTION

Several trends in the Information technology and business demands have called for development and implementation of more complex analytical techniques which go beyond the classical statistical analytical approaches. In the current information age there is an explosive expansion of digital data generated and stored in computer databases. Identifying potentially useful knowledge from such databases is not a trivial task and is resulting in the growing interest in data mining by researchers and practitioners. Another trend is the increasing pressure in companies to keep competitive advantage by construction and deployment of data-driven analysis. Data mining is a class of analytical techniques that examine a large amount of data to discover new and valuable information. It is a semi-automatic technique which can discover patterns, association rules, anomalies, and statistically significant structures in data. It has been increasingly recognized as a key tool for extracting meaningful information from the flood of digital data collected by businesses, government, and scientific agencies. Gartner Group, MIT Technology Review for 2001, and Palo

Alto Management Group have identified the data mining segment as one of the fastest growing in the Business Intelligence market and as a top ten emerging technology.

Data mining applications have proved highly effective in addressing many important business problems. Continuing construction and deployment of data mining for crucial business decision support systems has taken place. It has delivered measurable benefits: reduced cost, improved profitability, enhanced quality of service. Industries in which data mining became a necessary tool include insurance, direct-mail marketing, customer relationship management, fraud protection, network intrusion, telecommunications, retail, healthcare, science. The predictive modeling is often an integrated part of high-level of practical data mining.

This paper examines the specificity of image data mining, its challenges, and implementation model.

2. IMAGE MINING AREAS

There are several business channels for which the development of robust, scalable, and reliable image mining applications is critical: government surveillance and criminal investigation; science, including biomedical imaging and astrophysics images from telescopes and spectra data; health care and medical fields; art, design, and photographic art, and last, but not least, web mining for images. For example, the vast store of medical images leads naturally to using data mining techniques to discover interesting patterns and associations not previously known to physicians.

A comprehensive review of emerging scientific applications in data mining is given in [4]. The authors identify several scientific fields where a substantial growth of data mining is needed for further scientific advances. Biomedical engineering is in the midst of revolution with an unprecedented flood of data and images forcing biologists to rethink their approaches to scientific discoveries. Microscopic images of DNA, small molecular structures, gene representation in a population of cells – the availability of this data requires biologists to take opportunity with the automated learning abilities of data mining.

Geospatial data is another area where the scope and volume of digital geographic data sets acquired from satellites, high resolution remote sensing, and other monitoring devices easily overwhelm traditional spatial analysis techniques. Despite some prominent research and data mining algorithms developed in this area, geospatial data mining is in its infancy [9]. Climate and oceanographic data and Earth's ecosystem has in the recent years acquired copious amount of data through satellite images, terrestrial observations, and computer modeling. Periodical global snapshots, typically in a monthly basis provide a voluminous, large scale data that presents a major candidate for automatic rules and patterns extraction through image data mining, this complimenting textual data mining and traditional statistical techniques. Important results from effective data mining can be applied to identifying mineral, oil, and water resources, agriculture and forestry, land mapping, marine research, etc.

Data mining the large archives of digitized medical images of hospitals, healthcare institutions, and physicians' offices provides unique opportunity for advances in medicine and health care. MRI, PET scans, mammograms, ultrasound images, protein crystallography – all these image categories are excellent candidates for image mining [10]. Through image-content database query techniques, data mining can discover common and indicative patterns leading to early and automated discovery of abnormal organ tissues, lesions, cancer, and others. Moreover, it is also possible to use image-content data mining as a clinical tool for new cancer screenings as well as to provide clues to deeper understanding of the nature of different sicknesses and their correlation with environmental and genetic factors as derived from patient records.

Astronomy and Astrophysical Sciences also benefit tremendously from imaging and textual data mining applications. Hubble telescope and ground-base optical and radio telescopes supply astrophysicists with sky data tens of times more voluminous that any of the traditional image analysis and statistical techniques can handle. Lawrence Livermore National Laboratory in California, USA, is a leading provider of image mining applications tailored towards the tedious and arduous process of finding a handful of precious images of rare astrophysical objects or patterns by mining trillions bytes of image data. For example, their SAPPHIRE project (lead by Kamath, [6]) provided efficient search for a rare category of quasars among 22,000 sky images amounting to 100 gigabytes of image data.

One of the most urgent areas for applying advanced imaging data mining is the security and surveillance. The new government regulations, following September 11 terrorist attack, the national strategy of Homeland security and the new security regulations call for applying cutting edge imaging techniques including data mining. Human face recognition, signature recognition, automated target recognition and identification, X-ray image mining (for airports and facility security), and fingerprint/retinal recognition are estimated to be the fastest growing image mining application to respond to the heightened need of securing the population against criminal and terrorist acts.

3 IMAGE MINING PROCESS AND CHALLENGES

Image mining is the discovery of patterns from a collection of images [7]. The image mining is highly specific because the image databases are predominantly non-relational. In addition, many image attributes are not directly visible to the user. The size of each data item is large which imposes severe requirements not only on storage, but on data delivery (the query process, browsing results, etc.). The data mining implementation model proposed by [3] and verified in a case study by [5] is viewed as a framework of how data mining should be conducted. Several common stages in imaging mining can be identified [6]: Data Preprocessing, Pattern Recognition, and Knowledge discovery stage. These stages are iterative and highly interactive in nature. The whole process repeats until the useful knowledge is extracted.

3.1. Image Preprocessing

The image data is highly non-trivial. Preprocessing and postprocessing of image data are often the most critical, time consuming phase which determines the effectiveness of the data mining application. Data preprocessing is more influential and time consuming. In addition, it is domain specific which impedes the preprocessing automation. It is a well known argument in data mining that data and human issues are the bigger success or failure factor in data mining than the algorithmic and model issues. Some researchers report that data preparation phase is the most resource intensive stage – around 60% to 70% of the total effort in the data mining project is dedicated to preparation [3]. The Preprocessing stage includes De-noising of the images, Object identification, and Dimension reduction.

The sources of noisy data can be typographical errors, missing values, incorrect information, duplicate data, etc. In addition many images are not in a form suitable and have to be transformed to more meaningful attributes. To overcome the problem of noisy data, frequently a whole data cleaning system has to be implemented that reconciles the format differences by allowing the users to specify the mapping between attributes in different format styles and encoding schemes [1]. Some categories of data are more prone to noise than others. It depends on how controlled the environment can be at the time the image acquisition. For example, medical images, photographs taken in controlled conditions are much less noisy than telescope and remote sensing images where the data acquisition process is highly dependent on atmospheric disturbances. Various de-noising techniques include spatial filters, simple and wavelet thresholding, and other traditional image processing techniques for noise reduction.

The object identification is non-trivial. Among the greatest challenges is the tremendous variety of object shapes, hues, and contrasts, its image quality and boundaries. There is no universally accepted measure of quality of the object, in time sequence measures the object can change shape or move, split, or merge with another object. Among traditional image processing techniques, widely used are thresholding of the image histogram, segmentation techniques, and edge detection by using filters. Along with the object's identification, its features (distance parameters, angles, areas, etc.) are also extracted.

The Dimension reduction is made to facilitate the computational algorithms and shorten the time for pattern extraction. Classical methods of dimension reduction are applied, such as exploratory data analysis, principal and/or independent component analysis, etc.

3.2. Pattern Recognition

Traditional algorithms are employed at that stage. These include classification, clustering (segmentation), association/sequential pattern discovery, regression algorithms. A thorough review of the algorithms is

beyond the scope of this article. However, some studies report successful usage of specific algorithms in different application areas. For example, [8] used the fast Nearest Neighbor Search algorithm which can have broader applications in photo journalism and web image mining as well. (*"Find shapes similar to this"*). The reported techniques achieve classification based on shape and degree of change in shape over time (for tumors), by providing Time Evolution analysis and detecting correlation among shapes, diagnoses, and symptoms. More sophisticated image mining application incorporate numerical signatures of image features such as color, texture, size, and shape and temporal changes. Successful brain image mining was reported by using nonparametric regression [11].

Once patterns are discovered, it is difficult to distinguish the spurious from the significant ones. While the traditional statistical approaches offer the estimate of significance level, it is not possible to apply these approaches directly due to spatial and temporal autocorrelations.

3.3. Knowledge Discovery

Therefore, when genuine patterns are identified, domain-specific knowledge is essential so the patterns of no interest and/or relevance are filtered. Filtering the large volume of rules discovered needs a design of a separate pruning algorithm to remove insignificant rules. This algorithm has to be domain specific and reflect the common and relevant knowledge in the corresponding domain. For a given image database we proposed the following algorithm for knowledge discovery. First we construct a database with records containing the following structure: (imageID, C_1, C_2, ..., C_n, T_1, T_2, ..., T_m, S_1, S_2, ..., S_k, F_1, F_2, ..., F_l), where imageID is a unique identification of the image; C_1, C_2, ..., C_n, are the values of the color characteristics; T_1, T_2, ..., T_m, are the values of texture characteristics; S_1, S_2, ..., S_k, are the values of shape characteristics; F_1, F_2, ..., F_l are the high level semantic features, given by an expert in the field. The mining process is defined into two steps. First we find the frequent multidimensional value combinations and find the corresponding frequent features in the database. The combination of attribute values that occurs two or more than two times are called multidimensional pattern. For mining such pattern a modified BUC algorithm [2] is used. The second step includes mining the frequent features for each multidimensional pattern. They constitute the obtain rule base set for the high semantic features.

One of the most important aspects in this stage is the visualization, filtration, and validation of the newly discovered patterns and rules. This stage is highly domain and knowledge specific, and during the mining method validation it requires close interaction between the domain expert and the data mining researcher. The extent to which the knowledge discovery presentation is user-friendly and does not require additional training determines the success of the data mining application. Many end users' primary activities are not to analyze the outcome of decision support systems – for example physicians, astronomers, artists, photographers, etc. Therefore they cannot take the time to sort through large number of rules. It is therefore important to present the discovered rules in a visualized, easy to understand form. The data mining application should be able to assist the user in analyzing knowledge discovery quickly and easily.

4 CONCLUSIONS AND FUTURE RESEARCH

This overview of image data mining focuses on fundamental questions of image data mining implementation, usability and challenges. In addition to providing a discussion of usability in different business channels, implementation specificity and implementation model stages, the paper gives an overview of methodology used in the different phases of image data mining. Future research objective is to test and provide comparison of different data mining modeling techniques on astrophysical images provided from computer simulations of supernova explosions, as well as actual images obtained from telescopes. We intend to use about 500 images to extract and validate relevant rules and patterns by using the methodology described above. Additional area of interest are the privacy issues related to image data mining in health care and security and surveillance applications, as well as individual's vulnerabilities associated with the Internet availability of such image data.

5 REFERENCES

[1.] Apte, C., et al., 2002, Business Applications of Data Mining, *Communications of the ACM,* 45, 8, 49-53.

[2.] Beyer K., and Ramakrishnan, R., Bottom-Up Computation of Sparse and Iceberg CUBEs. SIGMOD'99.

[3.] Caben, P. et al., 1998, *Discovering Data Mining: From Concept to Implementation,* Prentice Hall, E. Cliffs, NJ.

[4.] Han, J. et al., 2002, Emerging Scientific Applications in Data Mining, *Communications of the ACM,* 45, 8, 54-58.

[5.] Hirji, K., 2002, Exploring Data Mining Implementation, *Communications of the ACM,* 44, 7, 87-92.

[6.] Kamath, C., 2001, The Sapphire Approach To Mining Science Data, http://www.llnl.gov/casc/sapphire

[7.] Kantardzic, M.. 2001, Data *Mining. Concepts, Models, Methods, and Algorithms*, Wiley Interscience.

[8.] Korn, F, et al., 1996, Fast Nearest Neighbor Search in Medical Image Databases. *Proceedngs of the 22nd VLDB Conference,* Bombay, India.

[9.] Miller, H. and Han, J., 2001, *Geographic Data Mining and Knowledge Discovery.* Taylor and Francis, London, UK.

[10.] Stanchev P., Fotouhi F., Siadat M-R., and Soltanian-Zadeh H., 2002, Medical *Multimedia and Multimodality Databases.* chapter 7 in Multimedia Mining. A High Way to Intelligent Multimedia Documents, D. Djeraba (Eds) - Kluwer Accademic Publishers.

[11.] Tsukimoto, H. and Morita, C., 1995, *Machine Intelligence,* pp. 427-449, Oxford University Press

(17.) Dinev T., Stanchev P., "Image Data Mining – Conceptual Overview of Implementation and Methodology Challenges", ***IADIS International Conference, e-Society 2003,*** Lisbon, Portugal , 3-6 June 2003.