"Alexandru Ioan Cuza" Iaşi University
Faculty of Economics and Business Administration
Doctoral School of Economics and Business Administration
Field of study: Business Information

Multimedia database applications – problems and solutions

PHD Thesis - Summary

Scientific coordinator:
PHD Professor Marin Fotache

Author:
PHD student Dan Andreea Irina

Iaşi, 2013

In this paper, the author wants to identify the unification possibilities of a relatively new field, of multimedia database technologies, with einvoicing methods, especially storing, archiving and retrieval of electronic documents. This paper wants to describe a possible structure of a retrieval and storing mechanism of the data on the invoices, identifying advantages and disadvantages of such an application.

The most of the paper is oriented on the technical part through image interogation algorithms and data retrieval information. The process of implementation and utilization of a storage and retrieval mechanism of electronic documents are also described. A number of case studies is included in the paper to verify the hypothesis. All these technical elements and components are based on theoretical sections that are destined to create a knowledge base. Using the literature in the field, the author describes the legal aspects and the einvoicing standardization. We try to present the actual view on the einvoicing market, in Europe, as well as in the rest of the world, to give a real view on the needs and existing problems.

In almost all the sections there are case studies that have the purpose of bringing in attention the practical frame of the studied phenomena, the statistic analysis of data sets, as well as experimental research, to simulate real situations that occur during the adimission process of the Faculty of Economy and Business Administration, in terms of validating the declared information on the enrollment papers and the data taken from the adimission folders. The conclusions of these studies

are generalized and possible applications in other practice fields are described.

This paper aims to present a possible solution to the problem of the increasing volume of data stored and processed without classifying and presenting the advantages and disadvantages of alternative solutions and wants to leave it up to the reader the decision whether the proposed solution is a viable one, with utility in electronic documents processing or other economic areas.

Key terms that form the basis of this work are the storage and processing. On the one hand, the management workflow of multimedia databases stores data in various forms, then processes through the various algorithms and auxiliary applications. On the other hand, electronic invoices require a minimum of archiving storage systems and data processing formats available on the images to be processed and sent to payers, all information verified and processed to integrate into the payers' ERP systems.

The problem of data storage and processing is influenced by the fact that there are produced increasingly more and more data. One of the main problems brought by this is hindering the processing, the response time for databases with a number of records that doubles every year gets bigger, which will affect the quality of results. In the image processing were formulated, in the last few decades, a number of algorithms for information retrieval, rather than using the metadata and the queries and return the results. Out of these, the most interest seems to be on the methods for data extraction in content, respectively by

comparing the pixels and comparison of images through the colors. The paper focuses on image processing, of all types of multimedia data as the most advanced stage of research is on this type of data, and because they can be used in practice. Many institutions scan and maintain data in image format to have backups of their documents. These copies can be used for many more uses than just protection in case of loss of original papers, they can be queried and compared with other documents of the same type and grouped them according to certain criteria, for easy finding.

The utility of electronic documents raises concern for the parties involved in e-invoicing, especially in recent years. Current legislation allows archiving for a specified period electronic format. For large companies that processes thousands of documents it is almost impossible to find a document in the stored set. A search algorithm based on key words, an usual practice in multimedia databases, could be the solution for such problems.

Not only the processing of electronic documents can use these methods. One of the case studies in this paper show the usefulness of multimedia databases in the college admissions. Admission documents from the files can be scanned and processed as multimedia data. Thus we have the features to eliminate a number of processing errors and to identify the types of documents and their validation. Also, other electronic documents, contracts, orders etc. can be stored as multimedia data, to be easily retrieved and processed. All these are just a few reasons that the

use of multimedia databases related to the scope does not stop in the economic field.

The purpose of this paper is to analyze the current situation of the market and technological ingredients of multimedia databases, and formulating new solutions to problems that require the use of multimedia databases, independently or in conjunction with other technologies such as OCR, regular expressions and so on.

In recent years, several software systems dealing with multimedia data processing have been presented in the literature as multimedia database systems. However, many of these systems would not be classified as a multimedia database system. On the other hand, many specialists support the multimedia data retrieval systems in the multimedia data collections, but they do not provide data or media independent efficiency and scalability. On the other hand, there are the expandable database of multimedia extensions, with insufficient support for the query, and inaccurate in terms of data recovery functionality provided in the first type of multimedia software systems. Computer scientists use the term ‚multimedia' to refer to something that is not a type of conventional alphanumeric data. Sometimes, the term becomes more explicit through a list of data types, using an intuitive notion of multimedia images, audio, video and text. Defining multimedia more accurate than using a list of types of media data is surprisingly difficult. Grosky, Fotouhi and Jiang tried to define multimedia by human activity involved in creating data. But long before the multimedia data was entered into databases, movies and music were created all the humans

with the specific care to communicate to the public, the message of the artist. The multimedia data information transmitted can represent anything from the real world, while traditional data information is a symbolic representation of the facts, limited to the universe of discourse of the database.

The multimedia data storage and of the normal data query that was derived from the multimedia data, does not change a database system in a multimedia database. A true the multimedia DBMS should not limit its users to access the predefined patterns. In a given service, one could assist developers specify component analysis queries that match a predefined conceptualization.

However, these conceptualizations predefined capture only the information necessary to respond to a subset of all possible questions about football videos, users should be able to formulate queries and inform contingency that they need (for example, the marketing department of some multinational companies might be interested in seeing how often was their banner in the stadium shown on television, during the last year). Of course, it may be impossible to build queries appropriate to the needs of complex, given the data digitized and automated analysis techniques known today. However, the decision whether to spend more time to try and find the relevant items for an ad-hoc query should be left to the user.

History of relational databases repeats itself with multimedia databases. Until recently, The multimedia applications included their multimedia logic, without using other systems or services. But now you can use

multimedia databases management systems, as the first applications that use relational database functionality developed more than three decades ago. Developing such systems is still in an early stage, so we can not yet speak of a standardized language for querying data. In most cases, the systems include their own language application stored and organized. This language, most often similar to SQL is adapted to the needs and requirements of the application.

Database applications are different from multimedia applications using traditional database where data are stored and queried. The multimedia data are different, complex, difficult to standardize in a common language. For example, audio data and video data are composed of other multimedia data. Aygun indicate that the multimedia objects are multidimensional and hierarchicaly structured. If we take, for example video files, they have taken behaviors from both audio and images and the time element plays an important role in image sequence and duration.

Four decades ago, IBM imagined identifying people using computers. IBM said then that this could be done by something the user knows or remember, or carries (access cards), personal or physical characteristic.

This concept has developed a new area of research: biometrics. Anil K. Jain, Patrick Flynn and Arun A. Ross start their biometrics textbook (A Handbook of biometrics) with the definition: "biometrics is the science of determining an individual's identity based on physical attributes, chemical or behavior of the person." Attention towards biometrics has fluctuated over the past decades as a result of variable attention to

security issues in society. In the 19th century the attention towards biometrics has been strengthened by the need for large-scale systems for managing attributes for identity recognition.

The widespread adoption of devices capable of capturing multimedia content in the form of pictures and videos created large volumes of information that cannot be indexed by traditional text-based files. The need for effective representation, indexing, searching and browsing multimedia content boosted research-based content extraction of data from multimedia systems (CBMR). CBMR systems are multimedia objects based on their visual and audio content, rather than textual descriptions by labels. The main limitations of existing representation methods include high dimensionality of feature descriptors and their inability to represent objects with homogeneous content. Unlike traditional methods that describe the content using simple statistics, regardless the homogeneity, the DD object approach identifies the optimal number of components to describe each object. Objects are represented by descriptors uniform properties with some components, while the homogeneous objects need a larger number of components. DD content description property offers vectors. It allows only comparing multimedia objects.

Multimedia representation focuses on efficient and accurate description of information that facilitates multimedia information retrieval. Different approaches have been proposed to map objects of multimedia data in formats that are easily processable, however, most of the proposed systems can not guarantee the accuracy when performing the

extraction of content. Therefore, the key issue in multimedia representation is a compromise between accuracy and efficiency.

Among the core issues mentioned above, multimedia representation, classification, and query processing provide the foundation for indexing. Adequate representation of multimedia entities have a significant impact on the effectiveness of multimedia indexing and retrieval of data. For example, the representation of object-level multimedia data usually provides more convenient ways than content-based indexing representation on pixel level. Similarly, queries are resolved by fields of multimedia data representation, either at the object level or pixel. Looking at the nearest neighbors scheme, they are usually based on careful analysis of multimedia representations of data and organizing content knowledge in multimedia systems.

Research in the image database systems have traditionally focused on the design of the robust processing techniques and image recognition. The growing role of images in multimedia applications has stimulated interest in the database administration files issues. Challenges in developing databases include processing images and extracting content dominan object identification data models to provide effective content-based indexing and indexing and fuzzy processing.

In terms of organizational structure, a multimedia document can be viewed as a collection of related information objects such as books, chapters, sections, etc. The logical structure of objects can be stored in the form of a meta scheme associated with each document. Meta information about these organizations can be used for searching and

accessing various parts of a document. Models of logical structure of multimedia documents can be independent of composition models. This independence can support different presentation styles for a document that can be adapted to the target audience and display constraints. Known organizational modeling paradigm is based on hypermedia documents. There are three types of links used in a hypermedia environment. These include links to the base structure for defining the organization documents associative links connecting concepts and access to the same information in different contexts, and referential links that provide additional information about a concept in a document.
In Chapter 1 of the thesis, the author describes in a practical study of the activity profile, dependencies in the choice of a electronic document processing standard, based on a description of the current market situation of e-invoicing in Europe and worldwide.

E-invoices are not a new invention. In the 60s, various organizations, particularly airlines, those that offer delivery services have realized that information processing speed is critical to remain a major player in the market. To avoid delays, errors and high costs to be reduced or eliminated paper documents. In 1979, the American National Standards Institute (ANSI) initiated the implementation of a standard between business partners, including the use of electronic invoices. As a result, Electronic Data Interchange (EDI) ANSI X-12 has determined that, in 1986, to establish United Nations Electronic Data Interchange for Administration, Commerce and Transport (UN / EDIFACT). Today, a great number of companies still using EDIFACT standards. EDI is the

true "paperless" method used in business processes such as procurement, ordering, delivery, paying bills, contracts, sales and more. However, the high cost of implementation methods based on the EDIFACT processing are needed for a cheaper alternative. In the early 2000s there were many attempts to find the cheapest solution. Among the solutions it was Electronic Invoice Presentment and Payment (EIPP), who used the Internet to send electronic invoices to the website of the supplier that the client can access. Although it was an acceptable solution providers EIPP hasn't been a real success, perhaps because they did not offer immediate solutions to reduce costs for customers. Despite these facts, EIPP is still used by many companies, especially in North America. Recently, Enterprise Resource Planning (ERP) platforms began to offer a solution for electronic invoice processing by computer systems. Many large companies have taken advantage of these features, scanning an impressive number of invoices (between 100.000 and several million annually). Many of these companies have realized that scanning is only a temporary solution, a step towards a fully electronic process. In an attempt to replace the scanning process, some companies have implemented point to point connections with their partners. This step led to the dissatisfaction of the parties, since only a small percentage of the total invoices procesed was electronic and holding both printed invoices and those electronics has led to confusion. All this led to the development of a service hub operator driven model, where a consolidator oversees thr traffic from supplier to customer. This solution proved to be the winner.

Both North America and Europe are available a large number of XML schemas. This is not a problem because the available systems support more than 25 different XML schemas. Differences exist in European countries, considering that Austria, Germany, Poland, Spain, Netherlands and the UK supports PDF documents. North America delivers more bills than all the European market, which is estimated at 34 billion invoices issued yearly. Limiting the adoption of electronic invoicing in the U.S. in particular is the habit of making payments using checks. Another significant difference is how markets have developed. While most corporations in Europe have started automating accounting systems, North America started with e-commerce. As a consequence, the expected growth is half the percentage of the increase in Europe. Looking in the AFC, the rate of increase is similar to that in Europe. Many Asian countries are in the process of approving the legislation for electronic invoices. There are countries in Asia where electronic invoices are still prohibited, such as China and Russia. Latin America is becoming a mature market development environment for invoices. Countries like Brazil, Chile and Mexico have implemented electronic invoice processing systems and actively encourage private sector companies to follow the government model which adopted e-invoicing software.

The paper continues in Chapter 1 with a case study applied to a sample of 50 companies that use the services of e-invoicing providers. The case study is based on an existing problem in many companies in various fields. A company decide to adopt an ERP system including processing

of invoices received and issued becomes a major problem. The main reason is that having a large number of business partners of various sizes, it is impossible to demand for them to adapt to their integration and processing of accounting documents needs. That means various formats, some less known, documents scanned and sent via email and a lot of other options reference. Also, all the big companies, which also have been the most receptive to the adoption of e-invoicing community as a form of processing invoices have partners with the same document processing practices, and small companies for which the investment in an electronic processing of invoices methos is not a depreciable investment in a short time. Even if an ERP system is purchased, it can process only one type of standardized formats, not all formats received or requested from suppliers. This is where often e-invoicing service providers interfere. They provide services for archiving, retrieval, processing, possibly checking, converting and sending documents. As we can see, the problem is the number of standards that such a system must convert into and from a given one. Most providers receive e-invoicing format and converts them into the format requested by customers. This paper focuses just on standards and formats preferred by large companies, investigate trends and preferences of companies in various fields.

It should be noted here that, for e-invoicing service providers the multitude of formats and standards accepted is not the only problem. When customers among companies that have partnerships in other countries or more difficult in countries outside the European Union

issues the documents. Each country has some legally binding information, while in others there is not the same requirement. As there are more exchanges of goods and services, with many regions, it becomes more difficult to compose a format that meets all without affecting the information transmitted. On the other hand, large companies have many subsidiaries, usually in different regions in the world, which affects the content of the documents. A subsidiary of Asia will most likely have other mandatory information than another one in Europe. The difficulty increases when both subsidiaries send to the same client, or receive from the same supplier. An important remark is that this work is just a case study applied to a sample of 50 companies, not a process of appointing a standard or another as the most viable or easily converted.

Chapter 2 of this paper describes the methods of querying multimedia metadata databases, algorithms, modules and external modules query similarity. For images based on text queries, metadata is used to use information on the colors, shapes, locations, etc.. At a higher conceptual level, the search may be used to identify such elements and features of the object sought. At the highest level, the name of activities, places or emotions can be used to generate a set of image results.

Keyword-based search is addressed to people who know exactly what they want to achieve, so they can formulate queries based on keyword matching. Even if you know the context, one does not necessarily know how to formulate the queries.

Descriptive text for the objects in the image are used. The data extraction indexed keywords for text search. Techniques of the semantic web ontology and metadata languages contributes to the defnition classes defined terminology with semantic metadata representation. After finding a point of interest, image, semantic ontology model, we can search for the connection between time and the image database. Not all relevant images are included in the result set. For example, imagines where the same person appears in different situations will not appear in the result set.

Searching for content-based image databases is theoretically similar with the search in relational databases, searching by the standard attributes, i.e. the primary key (i.e. the filename). Also, looking for images can be done using secondary characteristics (information such as date of issue, place of origin, author, etc.). If keyword search failed, then we make the search where there is no longer used a descriptor (e.g. Finding all images with a high proportion of green in the bottom of the picture). Search image content is based on the concept of derivation, meaning internal representation of the content of the image (pixels). This method has some disadvantages such as high costs and problems with manual indexing. For this type of search the method is using colors, textures or shapes visually represented. The characteristics of the images, which can be done by searching and identifying them are color that is identified by the color histogram, texture materials, the nature of the image segments and shapes (contours) and after the morphology represented in the pixels.

Forms can be represented through images, but also graphic vectors (e.g. polygons or general points sets). Patterns of similarity for polygons (2D) apply the approximation forms of sized rectangles. The similarity of the properties of the sides is calculated by partial similarity search using the Fourier transformation. Models of similarity for 3D object based on graphical vectors are represented by histograms based form approximation for surface segments using geometric hashing, the nearest point iterative or algebraic invariant moments.

In text-based image queries, the problem of incomplete annotations (Incomplete Annotation Problem - IAP) can affect the quality of queries. A standard method used TLS to solve this problem is pseudo relevance feedback (PRF) which updates user queries formulated by adding feedback terms selected automatically describing the most relevant results of previous queries. PRF assumes that the databases contain enough feedback to make a selection of the results. Otherwise, in IAP, we have short notes, which can generate complete results.

Jinming Min and Gareth JF Jones proposed a method to query images using an external source of information, namely Wikipedia. In the proposed method, the relevant items are identified first by title, comparing them with search terms in the query. The comparison is made using Jaccard coefficient, new results are combined with previous results of the same query to generate the final data set, the result of the query. Min and Jones argue that the results obtained by this method are much improved compared to standard PRF methods. Volume of online images has greatly increased in recent years, having, on one side the

advantage of having a larger database plain hard to apply queries, but also the risk of having irrelevant results is greater. Searching in a large volume of data mean text-based queries in most cases, and the content of the image (the pixel comparison). It is often difficult or unavailable looking after a pilot image, making the text-based search, a simpler solution, especially for precise annotation data collections. Added annotation are usually used to add image with the risk that the text is brief, irrelevant and subjective. By doing that an image could be the result of a query and can be ignored due to bad descriptions. In ad-hoc methods of data extraction (IR - Information Retrieval), a popular method to generalize the interogation is to formulate permisive criteria, a known method named query expansion (QE). This method gives the user the possibility to introduce elements that appear in documents and descriptive articles.

Going back, chapter 1 presents another case study, this time applied to a set of documents transmitted electronically. The case study is based on a real dataset, a sample of 5000 documents processed for companies in Europe in 2013. The data of this study is real, we removed all clues about the identity of the business partners. There are few studies that analyze real documents with values in documents, such as the total amount, applicable taxes, existing studies are generally made by companies seeking internal statistics on their workflows. In the last decade more companies in Europe have started to practice e-invoicing effectively without creating archives in paper format. Large volume of data has led companies to implement e-invoicing processes correctly.

Most providers of e-invoicing service not only provide processing and transmission of the documents, but their verification as well. Checking a document comes in addition to checking the sender and recipient documents and calculations contained in the document submitted. So, checking if the VAT is calculated in accordance with the law, if the date of delivery, date of the document are in accordance with the legislation in force. There are also various products and services for which different VAT rates are to be checked and compared. All checks affect each document submitted.

For this reason, it is interesting to analyze the number of occurring errors, the number of documents processed properly, VAT rates imposed and the degree of independence of variables in the data set. It is also interesting to analyze the frequency of document processing in a pair provider - customer and the number of correction documents, cancellation, sent over a period of time. A further objective of this study is to analyze the case of a document processing time from the time of its entry into the system by making the payment.

Chapter 3 of the paper presents an application for storing and processing images, useful in the verification of documents, such are diplomas. The data on documents is stored in PostgreSQL, and then placed in a system that extracts OCR text. The text is imported into an application of regular expressions to identify the resulted text, given certain fields, such as name, high school graduation, exam session, the grdes obtained during high school. The data found are imported into PostgreSQL database, and then compares the data entered manually

with the results the process described above. At the same time a document is checked by calculating the Euclidean distance and color histogram generation to confirm that the documents processed are really diplomas and transcripts or other documenta. Th purpose of the application is to eliminate the manual work done by admission committees to manually check all the data and provide a faster way to enter data into the databases of the faculty.

The author then presents their contributions to the field of study in the doctoral thesis. The first two chapters are theoretical, to give the reader a knowledge base to be applied in the following section. In the following sections the contributions are:

- Chapter 1:
    - Making a list of the storage needs of electronic invoices databases
    - Carrying out a statistical analysis on a sample of 50 of international companies with regards to the number of documents processed partners, etc.
    - A statistical analysis of a sample of 5,000 documents, identifying dependencies between certain attributes
- Chapter 2:
    - Writing a description for the representation and storage mechanisms in multimedia databases using metadata processing and clasification.

- A detailed description of several multimedia databases search algorithms and recommendations regarding each of the methods
- Creating a list of applications that use content based retrieval methods in their processes

- Chapter 3:
  - The implementation of a practical application for graduation diploma verification included in the admission file of the Faculty of Economics and Business Administration, extraction and pairing of the elements necessary to identify a candidate by using regular expressions and algorithms to compare the data obtained.

Future research directions are mainly directed towards the last part of the paper, namely practical application. This can be improved by optimizing response time for large volumes of documents and testing other types of documents, making one application used in other fields of economy. Also, another future research direction is to expand the case study applied to the sample of 50 companies, to a larger number and identifying additional variables that could influence the results of the analysis.