Computer Science Journal of Moldova, vol.**16**, no.2(47), 2008

New approaches for development, analyzing and security of multimedia archive of folklore objects

Galina Bogdanova, Todor Todorov, Tsvetanka Georgieva

Abstract

We present new approaches used in development of the demo version of a WEB based client/server system that contains an archival fund with folklore materials of the Folklore Institute at Bulgarian Academy of Sciences (BAS). Some new methods for image and text securing to embed watermarks in system data are presented. A digital watermark is a visible or perfectly invisible, identification code that is permanently embedded in the data and remains present within the data after any decryption process. We have also developed improved tools and algorithms for analyzing of the database too.

1 Introduction

Computerizing of science work in the folklore sphere is one of the newest modern problems, which should be comprehended not only as folklore texts collecting, but also in their wider context connection. The national collecting and conserving center for the Bulgarian folklore in the folklore Institute of BAS is the only special organization for documenting, conservation and popularizing the folklore in the country. It has a unique collection of authentic materials on different bases: paper documents, audio- and video cassettes, reel. Section 2 presents some basic challenges in digital organization of folklore materials. Section 3 is for the organization of a web based applications. In Section 4 we consider basics of relational database management systems. Section 5 is devoted to the basics of watermarking security. In Section 6 we present basics of

error-correcting codes. Section 7 considers the structure of the demo-version of Web based archive of folklore objects. Section 8 is devoted to the security of the system. In Section 9 is presented analyzing of multimedia archive of folklore objects. In Section 10 we investigate other folklore archives and compare them with the Web based archive of folklore objects considered in this paper.

2 Contemporary methods for digital organization of folklore archives

The idea of digital preservation of folklore texts and other folklore objects and their analysis with computer means, has been used since the early stages of such technique. Folklore explorers and programmers have already tried and succeeded in accomplishing this idea. Such pieces of work were made by S.E.Nikitina (Institute of Linguistics, Russian Academy of Science), Y.I.Smirnov (Institute of the World Literature Russian Academy of Sciences), the Russian folklore republican center and the folklore archives RGGU, IVGI RGGU etc. Some Web sites in the folklore area are [10], [9], [22], [26]. Actually computerizing of science work in the folklore sphere is one of the newest modern problems, which should be comprehended not only as folklore texts collecting, but also in their wider context connection.

There are two basic types of folklore information systems:

1) Data base text and electronic indexes;

2) Searching systems adapted to work with folklore objects.

The Database creation should be very precise as a method for folklore analysis with specific structure - irregularly, very branching with the "by hand" operating.

The national collecting and conserving center for the Bulgarian folklore in the folklore Institute of BAS contains examples in different spheres of traditional and contemporary culture: verbal folklore (fairy-tale and not fairy-tale prose, proverbs, sayings and etc.), rites

and festivity (calendar, family, labor etc.), musical, dance and plastic folklore (clothes, belongings, crafts etc.). Beside materials, collected from collaborators of the Folklore Institute, in the National Center terrain records of students' expeditions, as well as unique personal archives of famous folklore people and crafts are preserved. All country regions are available as well as the Bulgarian Diaspora in Czech, Slovakia, Hungary, Moldova, Ukraine. The regional genre makes the Center's collection very valuable and significant for our national culture. Today the archive funds of the Center are consistently filled up with folklore materials.

The development of calculating technique made collecting and preserving information easier. Great attention is paid to the protection of information from unauthorized access. During the last decade computer steganography has approved as a protective measure.

3 Organization of web-based applications

Most of the web-based applications now are using a multilayered architecture. It is organized as a two - tier structure. On the first level there is an application space. It consists of a web-browser which communicates with a web server via HTTP protocol. On the top level of the application space there are the server side applications. These are web servers, CGI scripts and API's for database connection. The second level of organization is the database tools. The most important application in this level is the database server used for storage and organization of the data in the system. Also we have tools and libraries for working with these database management systems. HTML (Hypertext Markup Language) is still the more important technology for visualizing in the Web. Although HTML evolved and many improvements have been added, it is in itself still static. The next step is the dynamic Web technologies, which allow the building of active web sites.

TECHNOLOGIES FOR BUILDING ACTIVE WEB SITES

These technologies could be classified as follows:

1) Dynamic technologies from the client side

   1.1) Java applets

   2.2) Active X controls

   3.3) DHTML (Dynamic HTML)

2) Dynamic technologies from the server side

   2.1) Common Gateway Interface (CGI)

   2.2) Active Server Pages (ASP)

   2.3) Java Servlets and Java Server Pages (JSP)

   2.4) PHP

   2.5) Patented API's for Web servers(ISAPI and NSAPI)

   2.6) Server Side JavaScript (SSJS)

More detailed explanation about all these technologies is given in [4]. We use DHTML and PHP in the development of demo version of web based client/server system with folklore objects. DHTML is not a standard defined by the World Wide Web Consortium (W3C). It is a combination of technologies used to create dynamic Web sites. DHTML allows to use regular HTML, scripts, document object model (DOM), absolute positioning, dynamic styles, multimedia filters and many other technologies for dynamic text and graphic manipulation, which HTML shows on the screen. Dynamic styles are based on Cascading Style Sheets (CSS). With CSS we have a style and layout model for HTML documents. CSS was a breakthrough in Web design because it allowed developers to control the style and layout of multiple Web pages all at once. To make a global change, we could change the style, and all elements in the Web are updated automatically. The HTML DOM defines a standard set of objects for HTML, and a standard way to access and manipulate HTML objects. Internet Explorer supports two languages for writing scripts - Visual Basic Script (VBScript) and JavaScipt. PHP stands for PHP: Hypertext Preprocessor. PHP is a server-side scripting language and works like ASP and JSP: the sections with scripts are inside the HTML page. It runs on different platforms (Windows,

Linux, Unix, etc.) and supports many databases (MySQL, Informix, Oracle, Sybase, Solid, PostgreSQL, Generic ODBC, etc.). As a WEB server for PHP scripts many different servers (Apache, IIS, etc.) can be used but we choose Apache HTTP Server. It is the only one that can use PHP as an internal module. This improves performance and security in the working process.

4 Relational databases and relational database management systems (RDBMS)

Database is a collection of records stored in a computer in a systematic way, so that a computer program can consult it to answer questions. There are a number of different ways of organizing the data in the database called modeling of the database structure. The model in most common use today is the relational model, which in layman's terms represents all information in the form of multiple related tables each consisting of rows and columns. Relational databases are combination of interconnected and stored in one place data with the presence of such minimal excess that allows their use in optimal way for one or more applications [29]. The fundamental assumption of the relational model is that all data are represented as mathematical n-ary relations, an n-ary relation being a subset of the Cartesian product of n domains. In the mathematical model, reasoning about such data is done in two-valued predicate logic, meaning there are two possible evaluations for each proposition: either true or false (and in particular no third value such as unknown, or not applicable, either of which are often associated with the concept of NULL). The relational model of data permits the database designer to create a consistent, logical representation of information. Consistency is achieved by including declared constraints in the database design, which is usually referred to as the logical schema. The theory includes a process of database normalization whereby a design with certain desirable properties can be selected from a set of logically equivalent alternatives. The access plans and other implementation and operation details are handled by the DBMS

engine, and are not reflected in the logical model. This contrasts with common practice for SQL DBMSs in which performance tuning often requires changes to the logical model. RDBMS are applications that are used to manage such a relational database. There are many such applications - Oracle, Sybase, PostgreSQL, MSSQL, MySQL etc. In our system we use Microsoft SQL Server for database management. The most important advantages of MSSQL server are:

1) RDBMS with high productivity

2) Developed to work with databases with complicated structure

3) Perform well in WEB environment - high speed, huge transactions

4) Full SQL compatibility - SQL is platform independent language for database manipulation.

SQL (commonly expanded to Structured Query Language) is the most popular computer language used to create, modify, retrieve and manipulate data from relational database management systems. The language has evolved beyond its original purpose to support object-relational database management systems.

5 Information protection with a digital watermark

The problem of information protection from unsanctioned access is solved already in antiquity. Even then two basic resolving directions have differentiated and now proceed to develop: cryptography and steganography [20]. Cryptography's designation is hiding a message by writing it in cipher, while in the steganography the mere fact for the existing of the secret message is hidden The word "steganography" has Greek origin, which means "secret writing" [14]. It is accomplished by several different methods and the similarity between them is that the secret message is put in a harmless, inconspicuous object. After that this object is transported openly to the address. In cryptography

the presence of a coded message itself attracts the attention. So it could be said that cryptography and steganography are not competitive information protection fields of study, on the contrary they can be used as self supplementing: one message can be encrypted and then sent by the secret steganographic methods.

Interest in digital watermarks has grown out of an increasing interest in intellectual property and copyright protection. Messages are hidden in digital data and especially multimedia: text, audio records, images, and video. A new steganography branch has appeared - digital steganography [5], [6], [7], [11], [13].

Digital steganography can be basically divided to four directions:

1) Embedding information for secret transferring;

2) Embedding a digital water mark (DWM) (watermarking);

3) Embedding identifying numbers(fingerprinting);

4) Embedding captions structure (captioning).

With embedding identifying numbers, every copy has its own number and that way, the further use of the product can be followed and the exact violator can be determined. Embedding captions is used for medical photos signing or laying an explanation legend on a map for example. The purpose is preserving of different information in a whole and stopping potential violators. The digital watermark is a special mark, imperceptibly embedded in an image, text or other signal in order to control its use. Embedding and retrieving of information from another is of basic importance in steganography and is done by the stegosystem's principles. The basic elements of a typical stegosystem for digital watermark are shown on Fig.1.

- precursory coder – structure for proper transforming of the secret message in order to embed it in the signal container (information sequence, in which the message is put);

- stegocoder – structure for embedding the secret message in other data and reading its specialties – structure for watermark retrieving;

- stegodetector – structure for stegomessage's presence determination;

- decoder – structure for secret message' s restoring.

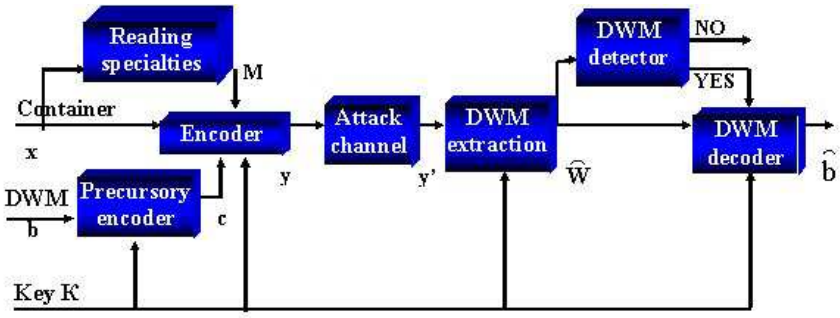

Figure 1. Stegosystem for digital watermark

Before watermark embedding appropriate transformations are necessary so that it corresponds to the container. For example, if the container is an image, then the watermark must be a two-dimensional array of bits. All transformations are done by the precursory encoder. Calculation of the general Fourier transformations for the message and the container are done in it. That enables embedding in the spectral area and that increases the stability of the watermark. An embedding key (K) is often used to increase the secrecy. Embedding and transforming messages in the container are done by the encoder.

There are different methods for that, which depend on the container's character and will be referred later. There are detectors for finding an existing watermark and for selecting it. In the first case detectors are possible with either hard or soft resolve. Metrics as Hamming distance and mutual correlation between the initial and delivered signal are used for choosing the proper resolve. When the initial signal is unknown, statistic methods are used. There are three types

of stegosystems depending on the necessary information for detection: open, half-hidden and hidden.

The hidden type I stegosystems have the best stability against outer intervention. To be stable a stegosystem must follow the next requirements:

- The system security must be completely determined by the secrecy key, so that even if the violator knows all algorithms of the system, he can not get additional information if there is a message in the container or not;
- The knowledge for the presence of a message in a container must not give the violator an opportunity to discover such in another container;
- Putting the message must not influence the container's quality;
- Possibility for discovering a secret message where it does not exist must go to zero;
- To have considerably complex algorithms for coding and decoding.

Necessary requirements for the DWM:

- The DWM must easily recover the legal owner;
- It must be steady to:
  - o common signal processing;
  - o common geometric transformations – rotation, translation, scaling and others;
  - o collusion and forgery.
- Universality – one and the same watermark for all types of multimedia;
- Unambiguous.

There are three types of stegosystems according to their steadiness to outer influence: steady, fragile and semifragile. Fragile DWM are destroyed only by insignificant modification of the container. They are used for signal authentification. The difference from the digital sign is that fragile DWM allow little modifications, which is important for the compression for example. Semifragile DWM are steady to some influences and fragile to others. As a whole all DWM can assign to this stegosystems type. For example, they can allow compression but they are fragile to elements removal.

5.1 Methods for image watermarking

5.1.1 Methods for hiding the data in the spatial area

The data is put directly into the initial image. Its basic advantage is that no additional transformations of the image are needed. DWM is embedded by manipulating the brightness or the component colors. Its basic disadvantage is the weaker steadiness. Such a method, used in the protection of the concerned folklore system, is Kutter's method [15]. Let s be a single bit to be embedded in an image $I = (R, G, B)$, and $p = (i, j)$ a pseudo-random position within I. This position depends on a secret key K, which is used as a seed to the pseudo-random number generator. The bit s is embedded by modifying the blue channel B at position p by a fraction of the luminance $L = 0,299R+0,587G+0,114B$ as: $B_{ij} = B_{ij} + (2s - 1)L_{ijq}$ where q is a constant determining the signature strength. The value q is selected such as to offer best trade-off between robustness and invisibility. In order to recover the embedded bit, a prediction of the original value of the pixel containing the information is needed. This prediction is based on a linear combination of pixel values in a neighborhood around p. The sign of the difference between the prediction and the actual value of the pixel determines the value of the embedded bit.

$$B'_{ij} = \frac{1}{4c} \sum_{k=-c}^{c} B_{i+k,j} + \sum_{k=-c}^{c} B_{i,j+k} - 2B_{ij}$$

where c is the size of the cross-shaped neighborhood. To retrieve the embedded bit the difference δ between the prediction and the actual value of the pixel is taken: $\delta = B_{ij} - B'_{ij}$. The sign of the difference δ determines the value of the embedded bit. Retrieving of the bit is made without the knowledge of an initial message. For that purpose prediction of its initial value is made according to the value of the near pixels. Accuracy can not be always guaranteed in the prediction of the secret bit value, so the embedding and retrieving functions are not mutually convertible. Some additional techniques, concerned in, are used. Also, robustness could be improved with the use of optimal error correcting codes. The method is steady to filtering, JPEG compression, geometrical transforms.

5.1.2 Spread spectrum watermarking

Based on different transformations of the container, for example Discrete Cosine Transform (DCT), it gives opportunity for greater steadiness of the DWM to transformations [7].

5.2 Method for text watermarking

5.2.1 Line-Shift Coding

This is a method, in which text lines are shifted vertically, so that the document could be uniquely coded. {-1, 1, 0} is often used as an initial alphabet, which is coded as shifting the line up, down or leaving it to its place. In most cases decoding can be made without the usage of the original document, if the constant space between the lines is known. It is easily found but steady to noise. This method is used for protection of .doc .ps in the folklore system.

5.2.2 Word-Shift Coding

This is a method of altering a document by horizontally shifting the locations of words within text lines to encode the document uniquely. The method is least visible when applied to documents with variable

spacing between adjacent words. Because of this variable spacing, decoding requires the original image – or more specifically, the spacing between words in the unencoded document.

5.2.3 Feature Coding

The image is examined for chosen text features, and those features are altered, or not altered, depending on the codeword. Decoding requires the initial document. The choice of text characteristics can be made by different criteria [6].

6 Error-correcting codes

6.1 Basics

The object of an error-correcting code is to encode the data, by adding a certain amount of redundancy to the message, so that the original message can be recovered if not too many errors have occurred [12].

Definition 1 *A* $q-$ ary code *is a given set of sequences of symbols where each symbol is chosen from a set* $F_q = \{\lambda_1, \lambda_2, \ldots, \lambda_q\}$ *of* q *distinct elements.*

The set F_q is called the alphabet and is often taken to be the set $Z_q = \{0, 1, 2, ..., q-1\}$. If q is a prime power we often take the alphabet F_q to be the finite field of order q.

Definition 2 *A binary code is a given set of sequences of 0s and 1s which are called codewords.*

Definition 3 *The (Hamming) distance between two vectors* x *and* y *of* $(F_q)^n$ *is the number of places in which they differ. It is denoted by* $d(x, y)$.

Definition 4 *Let* F_q *is the Galois field* GF(q), *where* q *is a prime power, and let* $(F_q)^n$ *is the vector space* $V(n, q)$. *A linear code* C *over* GF(q) *is a subspace of* $V(n, q)$, *for some positive integer* n.

If C is a $k-$dimensional subspace of $V(n,q)$, then it is called $(n,k,d)-$code, where n is length, k is dimension and d is the minimum distance of the code. Sometimes we denote it just (n, k) code.

Definition 5 *We call an* $(n,k,d)-$code *optimal if for fixed n, k it has the largest possible d.*

6.2 Reed - Solomon codes

Reed - Solomon (RS) codes are nonbinary cyclic codes with symbols made up of m-bit sequences, where m is any positive integer having a value greater than 2. RS(n, k) codes of m-bit symbols exist for all n and k for which

$$0 < k < n < 2^m + 2$$

where k is the number of data symbols being encoded, and n is the total number of code symbols in the encoded block.

Now let's make a more precise definition of RS codes [21]. Let α be a primitive element in GF(2^m). This means that α is an element of GF(2^m) such that each nonzero element of the field can be represented by a power of α. In these conditions for any positive integer $t \leqslant 2^m - 1$, there exists a t-symbol error-correcting RS code with symbols from GF(2^m) and the following parameters:

$$n = 2^m - 1 n - k = 2tk = 2^m - 1 - 2td = 2t + 1 = n - k + 1$$

The generating polynomial for an RS code takes the following form:

$$g(X) = g_0 + g_1 X + g_2 X^2 + \ldots + g_{2t-1} X^{2t-1} + X^{2t}$$

where $g_i \in$ GF(2^m) and $g(x)$ has $\alpha, \alpha^2, \ldots, \alpha^{2t}$ as roots.

One of the most important features of RS codes is that the minimum distance of an RS(n, k) is $n - k + 1$. Codes of this kind are called "maximum distance separable codes" (MDS). RS codes achieve the largest possible code minimum distance for any linear code with the same encoder input and output block lengths.

Also Reed-Solomon codes have an erasure-correcting capability, ρ , which is:

$$\rho = d - 1 = n - k$$

Simultaneous error-correction capability can be expressed as follows:

$$2\alpha + \gamma < d < n - k$$

where α is the number of symbol-error patterns that can be corrected and γ is the number of symbol erasure patterns that can be corrected.

There are many proposed algorithms for effective encoding and decoding of RS codes [21].

7 Structure of the application

The computer folklore system we are presenting consists of WEB interface for folklore database management. Client-server technology is used with the browser on the client computer (MS Internet Explorer has the best effectiveness), and for server Apache HTTP Server. Program technologies on the client side are ActiveX and Jscript, and on the server side – PHP scripts. The database is of relational type and is managed by RDBMS (System for database management) of MSSQL server.

The relations and the type of tables organization in the database are shown in Figure 2, described in detail in [17].

It is an open system, which is specified, changed and renovated in the process of categorization with the rising of the processed archive documents. The scheme is based on several important principles – presenting the folklore culture in its diversity and orderliness, pointing out that every culture fact is unique, ensuring efficient access and easy use of the archive funds for everyone who is interested. The classification scheme is a projection of the basic culture spheres – songs, rites, speech, music, dance, household goods – more and more independent scheme

structures are formed in these main spheres. Without the inculcation of a particular scheme categorization is impossible. Its main purpose is fast and beneficial orientation in the enormous diverse source of folklore materials. It is also a necessary step towards modern processing of the empiric data. On the scheme from Figure 2 the database tables used in the system are shown.

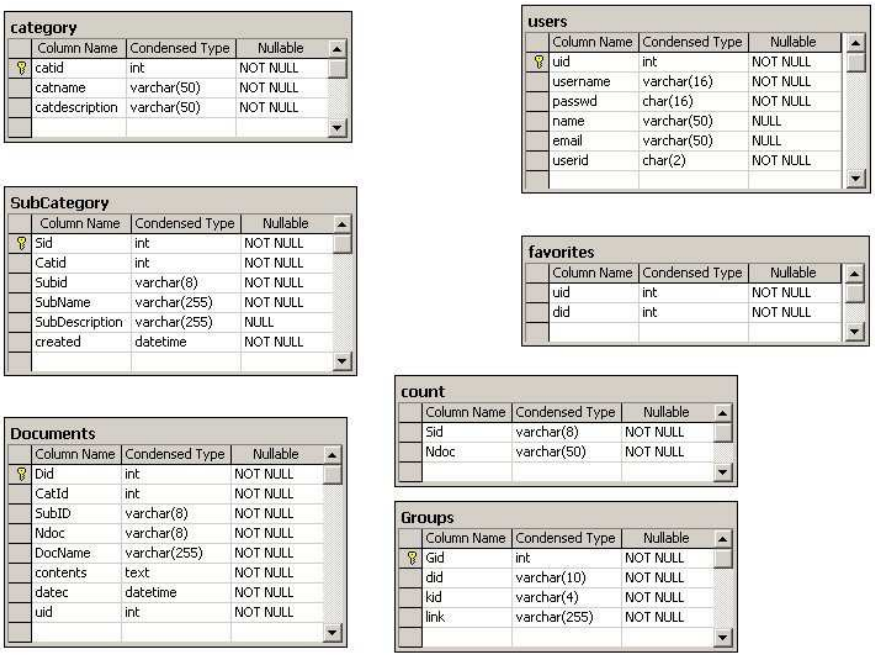

Figure 2. Database organization

- Category – information about the basic folklore categories
- SubCategory – information about the basic folklore sub-categories
- Documents – store of all folklore documents

- Users – information about all registered users
- Favorites – reference to all documents that are stored in "My Documents" category
- Count – Number of the documents in given category
- Groups – Used for documents additional grouping and external linking

The data is divided into three structural parts – audio, video and text archive and the mutual relations between the folklore parts are under review. The data is organized to allow flexibility in searching, renewing and adding.

8 Building the security of the data in the demo version of the web-based database for folklore documentation

In view of the fact that unique photo material is kept in the system, protection against illegal copying and spreading is needed. There is an opportunity for downloading partial photo materials on the client computer, but before that, it is uniquely signed, so that its origin could be proved. This is made by a specially developed ActiveX control, written in Borland Delphi and only the administrator has access to the system. It takes the graphic file as an entrance, which will be protected by the techniques described in 2.1.1, and puts a unique watermark. There is an opportunity for identifying the signed image. In 2.1.1 a primitive error correcting code based on the multiplicity of the embedding is used. It is well known that redundancy codes are far from optimality. So here we use some more effective error-correcting schemes. First we use hamming codes. They provide a mechanism that can be inexpensively implemented. In general, their use allows the correction of single bit errors per unit data, called a code word. Here we use a (15, 11) Hamming code that enables data link packets to be constructed easily by

permitting one parity byte to serve two data bytes. For error detection we use CRC. Second approach that we apply for error-correction are Read-Solomon codes [2]. RS codes are often used as "outer codes" in a system that uses a simpler "inner code". The inner code gets the error rate down and the RS code is then applied to correct the rest of the errors. Let RS(n, k) is a code over GF(2^m). Every element in this field can be represented uniquely by a binary m-tuple, called m-bit byte. To encode binary data with such a code a message of km bits is first divided into k m-bit bytes. Each m-bit byte is regarded as a symbol in GF(2^m). The k-byte message is then encoded into n-byte codeword based on the RS encoding rule. After the RS code is selected for the given case we proceed with the selection of the "inner code". According to the value of m we have selected for the RS code the same value should be selected for the dimension of the inner code. This code will correct errors on bit level in each of the m-bit bytes. Series of macros are developed; they use 2.2.1 to protect MS Word documents before they are given to the user. An additional password is also set, to prohibit him from the opportunity to change the document and that way to impede the removal of the watermark.

The data is divided into three structural parts – audio, video and text archive and the mutual relations between the folklore parts are under review. The data is organized to allow flexibility in searching, renewing and adding. For the security preventing unauthorized access to the data access levels are developed, with usernames, passwords and appropriate rights for them. We have three user roles – user, registrant and administrator. Here are the permissions that the different roles have:

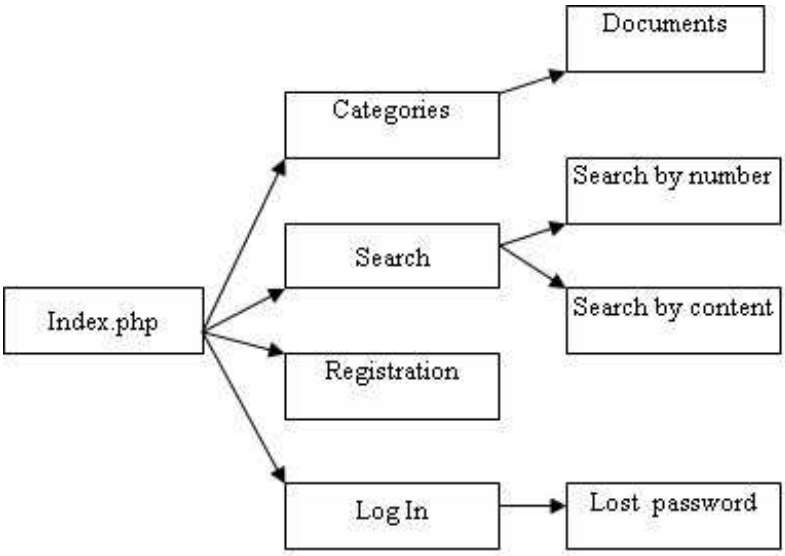

Figure 3. User permissions

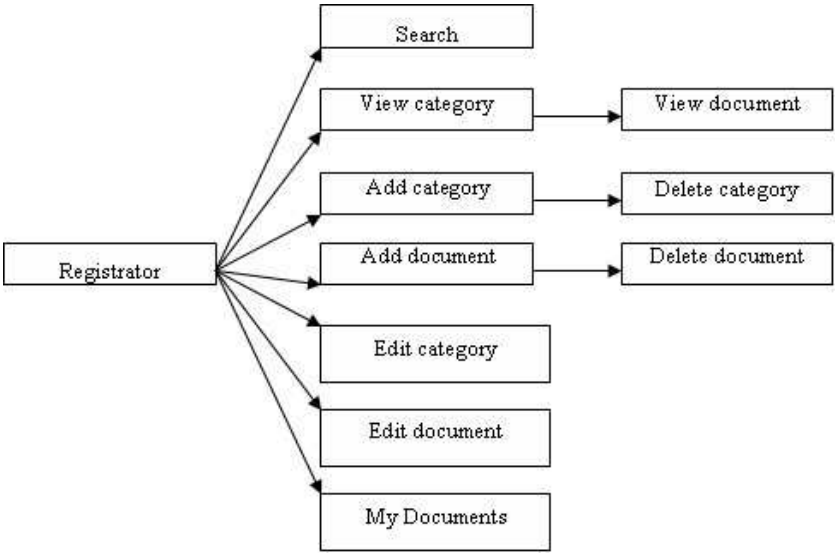

Figure 4. Registrant permissions

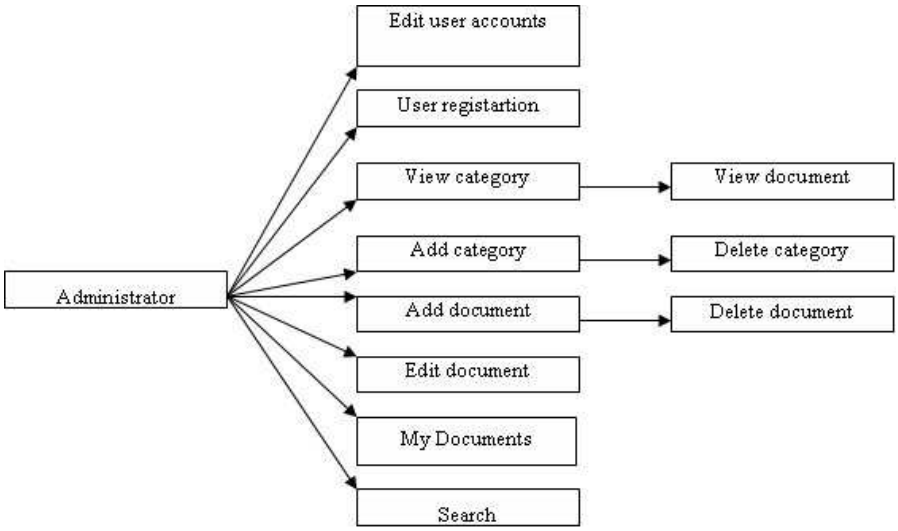

Figure 5. Administrator permissions

9 Analyzing the data of multimedia archive of folklore objects

Online Transaction Processing (or OLTP) is a class of programs that facilitate and manage transaction-oriented applications, typically for data entry and retrieval transaction processing. Some applications of OLTP include electronic banking, order processing, employee time clock systems, e-commerce, and eTrading. Online Transaction Processing has two key benefits: simplicity and efficiency. Reduced paper trails and the faster, more accurate forecasts for revenues and expenses are both examples of how OLTP makes things simpler for businesses. It also provides a concrete foundation for a stable organization because of the timely updating. OLAP is an acronym for On Line Analytical Processing. It is an approach to quickly provide the answer to analytical queries that are dimensional in nature. It is part of the broader category business intelligence, which also includes Extract transform load (ETL), relational reporting and data mining. Databases config-

ured for OLAP employ a multidimensional data model, allowing for complex analytical and ad-hoc queries with a rapid execution time.

The database in data warehouse is designed and the data is extracted from the OLTP (online transaction processing) database, transformed to match the data warehouse schema, and loaded into data warehouse database periodically by execution a batch job.

The data cube FolkloreCube is created in correspondence with the star schema of the dimensional model of the database in the data warehouse (fig. 6).

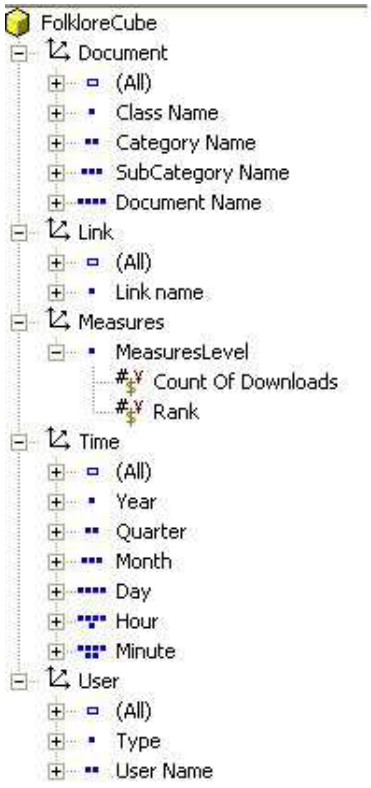

Figure 6. Folklore cube

It consists of four dimensions – Document, Link, User and Time. The first measure of the examined data cube is count of downloads of the folklore materials from the documents by the users. The users can rank the materials with the integer values between 1 and 7 that are stored in the measure rank and reflect the preferences of the users.

Applying the OLAP (Online Analytical Processing) Operations

MDX (multidimensional expressions) queries are applied to the data cube FolkloreCube providing the dimensional view of summarized data. Additional statements of MDX queries and the results from their executions are represented in [3], [8].

Discovering the Association Rules

The application for discovering the association rules in data cube FolkloreCube by using the OLAP operations is developed. An association rule shows the frequently occurring patterns (or relationships) of a set of data items in a database. An association rule is an implication of the form $X \rightarrow Y$, where $X = \{x_1, \ldots, x_m\}$ and $Y = \{y_1, \ldots, y_n\}$ are sets of items with $X \bigcap Y = \emptyset$. The rule $X \rightarrow Y$ has support s if $s\%$ of all itemsets contain $X \bigcup Y$. The rule $X \rightarrow Y$ has confidence c if $c\%$ of itemsets that contain X also contain Y. For example the following rule is generated from the data cube with daily downloads of folklore materials: {Document("*Songs*"), Time("*11/2006*")} $\rightarrow$ {Link("*somelink11*")} with support s = 0.1151 and confidence c = 0.2667. This rule means that one of the most downloaded materials from the documents in the category "*Songs*" during November 2004 is the material "*somelink11*" (with 26.67% confidence) and such downloads represent 11.51% from all downloads under study.

Discovering the Distribution Intervals of the Association Rules

In addition to the association rules the important practical applicability has their distribution in time. An algorithm that applies the OLAP operations and uses the fractal dimension to uncover the behavior of the association rules in time dimension is described. The code, which realizes the proposed algorithm, is successfully implemented.

Association rule mining aims to extract interesting correlations, fre-

quent patterns, associations or casual structures among sets of items in the transaction databases or other data repositories [1]. We use the proposed algorithm and software implementation to explore the behavior of the database in the multimedia archive of folklore objects. Association rules are widely used in various areas such as telecommunication networks, market and risk management, inventory control etc., where the proposed software tools could be successfully used.

10 Comparison with existing digital archives

We make an investigation to some other folklore software systems all over the world and compare them with the one described in this paper. According to [28] and to our research of other similar systems, we can conclude that two of the main criteria for comparison are its multimedia attributes and data security.

First we make a comparison by some attributes of a system that determine it as a multimedia:

- Text
- Audio
- Graphics
- Video
- Search

Results are displayed in the Table 1.

We could conclude that there are just a few archives that meet all the criteria to be called "multimedia". Also all the archives are not very flexible in their organization and are developed just to satisfy the needs of the particular folklore organization. On the other hand the system presented in this paper has a very flexible data organization that could be easily adopted to the needs of many different folklore organizations. It has all the aspects of multimedia archives too.

Table 1. Comparison of multimedia folklore archives

System	Text	Audio	Graphics	Video	Search
[16]	+	-	+	+	+
[18]	+	+	-	+	+
[9]	+	+	+	N/A	+
[24]	+	+	+	N/A	+
[27]	+	+	+	+	+
[25]	+	-	-	-	+
[22]	+	+	+	+	-
[23]	+	+	+	+	N/A
[19]	+	+	+	+	+
[26]	+	-	-	-	-

Second part of comparison includes data security and data analysis. In fact all the systems that we research include data security just in permissions level. The newly presented system also has different user roles with different access level. But more powerful tool in it is the newly proposed watermarking scheme and the software tools for watermarking embedding. They preserve data ownership even when it's already downloaded. Also the unique for this system, that we can't find in any other system are the data analyzing tools included in it. They allow more easy and flexible research of the data in the system and to analyze working activity of the different users.

11 Conclusion

We present new approaches used in development of the demo version of a WEB based client/server system that contains an archival fund with folklore materials of the Folklore Institute at Bulgarian Academy of Sciences (BAS). We use two error-correcting schemes – one that uses Hamming codes and another that combines Reed-Solomon codes as outer code and optimal linear code as inner code. We use them as a

part of watermarking software in order to secure data in the computer system. Finally we present improved software used to analyze the database. We could conclude that this is a useful and unique computer system that has very powerful tools for data security and analysis.

References

[1] Agrawal, R., Imielinski, T., Swami, A., *Mining Association Rules between Sets of Items in Large Databases*, In Proceedings of the ACM SIGMOD International Conference on Management of Data, Washington, 1993, pages 207–216.

[2] Berger, T., Todorov, T., *Improving the watermarking process with usage of block error-correcting codes*, Serdica Journal of Computing, 2008(submitted).

[3] Bogdanova, G., Georgieva, Ts., *Finding the error-correcting functional dependency by using the fractal dimension*, In Proceedings of the the fourth international workshop on optimal codes and related topics, 2005, 20–26.

[4] Bogdanova, G., Todorov, T., Blagoev, D. and Todorova, M., *Use of Dynamic Technologies for WEB-enabled Database Management Systems*, International Conference on Computer Systems and Technologies (CompSysTech'2003), Sofia 2003, II.22-1-6.

[5] Brassil, J., Low, S., Maxemchuk, N., O'Gorman, L., *Electronic marking and identification techniques to discourage document copying*, Proceedings of IEEE INFOCOM '94, 1994 3, 1278-1287.

[6] Brassil, J., Low, S., Maxemchuk, N., O'Gorman, L., *Hiding information in document images*, Proceedings of the 29th Annual Conference on Information Sciences and Systems, 1995, 482–489.

[7] Cox, I., Kilian, J., Leighton, T., Shamoon, T., *Secure spread spectrum watermarking for multimedia*, Proceedings of the IEEE International Conference on Image Processing. Vol. 6. , 1997, 1673–1687.

[8] Demetrovic, J., Katona, G., Miklos, D., *Functional dependencies distorted by errors*, Discrete Mathematics(accepted).

[9] Fife Folklore Archives, *http://library.usu.edu/Folklo/*.

[10] Folklore Databases, *http://www.eastern.edu/library/www/webindex/arts/folklore.shtml*.

[11] Gribunin, G., Okov, I., Turincev, I., *Cifrovaia steganographia*, Solon-Press, 2002 (in Russian).

[12] Hill, R., *A first course in coding theory*, Calendar Press, Oxford, 1986.

[13] Karasev, A., *Komputernaia tainopis grafika i zvuk priobretaut podtekst*, Mir PK. - 1/97, 132–134 (in Russian).

[14] Katzenbeisser, K., Petitcolas, F., *Information hiding techniques for steganography and digital watermarking*, Artech House Books, 2000.

[15] Kutter, M., Jordan, F., Bossen, F.,*Digital signature of color images using amplitude modulation*, Proc. of the SPIE Storage and Retrieval for Image and Video Databases. V. 1997. Vol. 3022., 518–526.

[16] Living Treasures, *http://www.treasures.eubcc.bg/* .

[17] Mateeva, V., Stanoeva, I., *Klasifikacionna shema na tipologichnia katalog v Instituta za folklor*, Bulgarski folklor. kn.2-3, 2001, 96–109 (in Bulgarian).

[18] Music Multimedia Archive, *http://musicart.imbm.bas.bg/en/about.htm*.

[19] Philadelphia folklore project, *http://www.folkloreproject.org/*.

[20] Privacy Guide: Steganography, *http://www.all-nettools.com/privacy/stegano.htm*.

[21] Skallar, B., *Digital Communications: Fundamentals and Applications*, Prentice-Hall, 2001.

[22] The American Folklore Center, *http://www.loc.gov/folklife/index.html*.

[23] The Estonian Folklore Archives, *http://www.folklore.ee/rl/era/*.

[24] The Folklore Program at the University of California, Berkeley, *http://ls.berkeley.edu/dept/folklore/*.

[25] The Israel Folktale Archives (IFA), *http://www.folklore.org.il/asai.html.*

[26] The Site for American Folklore, *http://www.americanfolklore.net.*

[27] The Ukrainian Folklore Archives, *http://129.128.116.48:8890/photo_ archives/.*

[28] Velvheva, J., Petkov A., *Informacionni sistemi i tehnologii v biznesa*, Russe University press, 2002 (in Bulgarian).

[29] Welling, L., Thomson, L., *PHP and My-SQL WEB development*, Sams Publishing, 2003.

Galina Bogdanova, Todor Todorov, Tsvetanka Georgieva Received June 4, 2008

Galina Bogdanova
Institute of Mathematics and Informatics
Bulgarian Academy of Sciences
Section Mathematical Foundation of Informatics
P.O.Box 323
5000 V.Tarnovo, Bulgaria
E–mail: *galina@moi.math.bas.bg*

Todor Todorov
Institute of Mathematics and Informatics
Bulgarian Academy of Sciences
Section Mathematical Foundation of Informatics
P.O.Box 323
5000 V.Tarnovo, Bulgaria
E–mail: *todor@moi.math.bas.bg*

Tsvetanka Georgieva
University of Veliko Tarnovo
Department of Information Technologies
3 G. Kozarev str. 5000 Veliko Tarnovo, Bulgaria
E–mail: *cv.georgieva@uni − vt.bg*