Digital Information Retrieval

Daniel Volovici[1], Macarie Breazu[2] Adi-Cristina Mitea[3], Daniel Ionel Morariu[4]

"Lucian Blaga" University of Sibiu, 10, Victoriei Blv., 550024, Sibiu, Romania

[1]daniel.volovici@ulbsibiu.ro; [2]macarie.breazu@ulbsibiu.ro; [3]adi.mitea@ulbsibiu.ro;
[4]daniel.morariu@ulbsibiu.ro

Abstract: The retrieval of digital information imposes the need for a special attention to the digital representation of documents. Data reduction, based on informational measures, represents the theoretical base for the optimization of digital documents storage. In this paper, besides analyzing data reduction and documents storage, we have also studied the multimedia data models related to the MMDBMS architecture. These theoretical aspects have been applied to the development of the ***SCRIBe*** – Information System for Processing and Visualization of Old Books Inventory.

Keywords: Information Theory, Document Representation, Information Retrieval, Digital Libraries.

1. Introduction

Information Theory [3] answers two fundamental questions in communication theory: which is the best compression of data (entropy H) and which is the best communication rate for data transmission (channel capacity C)? Therefore, some consider information theory as a subfield of communication theory. Indeed, it has fundamental contribution to statistical physics (thermodynamics), computer science (Kolmogorov complexity and algorithm complexity), statistical inferences and probabilistic theory and statistics.

Shannon claims that random processes as music and voice have an irreducible complexity the signal cannot be compressed below. He calls this Entropy, by association with its use in thermodynamics.

Information Theory proposes the means to obtain the extreme limits of communication. Although, these theoretical optimal communication schemes, even if wonderful, prove not to be computational feasible. Advances in the field of integrated circuits and design of codes allow us to obtain some of the gains suggested by Shannon theory. A good example for applying these ideas from information theory is the use of error correcting codes on CDs.

Modern research on aspects of information theory in communication focused on information theory in networking consisting of the theory of simultaneous communication rates from multiple senders to multiple receivers in a network. Some gains in the communication rates between the senders and the receivers were unexpected, but have a certain mathematical simplicity. Although, a unifying theory still has to be found.

Computer science (Kolmogorov complexity). Kolmogorov, Chaitin şi Solomonoff suggested the idea that the complexity of a string of symbols can be defined as the length of the shortest binary program that generates that string. Then, the complexity is given by the length of the minimal description. This complexity definition is universal, independent from computer, and of a fundamental importance. Kolmogorov complexity sustains the descriptive theory of complexity. Fortunately, Kolmogorov complexity approximately equals Shannon entropy H, if the sequence is chosen at random from a distribution having the entropy H. We consider Kolmogorov complexity to be more fundamental than Shannon entropy. It is the limit of data compression and leads us to a logical consistent inference procedure.

A pleasant complementary relationship between algorithmic complexity and computational complexity exists here. We can see computational complexity (time complexity) and Kolmogorov complexity (the length of the program or descriptive complexity) as two axes corresponding to the running time of the program and the length of the program. Kolmogorov complexity focuses on minimizing along the second axis and computational complexity focuses on minimizing along the first axis. Few attempts have been made to minimize simultaneous along both axes.

Physics (thermodynamics). Statistical mechanics is the place of birth for entropy

and the second law of thermodynamics. Entropy always grows.

Mathematics (Theory of probability and statistics). Fundamental measures of information theory – relative entropy and mutual information – describe the behavior of long sequences of random variables and allow us to estimate the probability of rare events and to find the best error estimator in hypothesis testing.

Philosophy of Science (Occam Razor). William of Occam said: „Causes shall not be multiplied beyond necessity" or, paraphrasing him, "The simplest explanation is the best". Solomonoff, and later Chaitin, stated that we can obtain a universal good prediction procedure if we take a weighted combination of all the programs that explains the data and still look at what they print afterwards. Furthermore, this inference procedure will work in many problems that cannot be solved by statistics. When it is applied on stock market is has to find stock market rules and to extrapolate them optimally. Basically, such a procedure would have found Newton law in physics. Certainly, such inference is not applicable, because eliminating all the programs that do not manage to generate the data takes too much time.

Economy. Repeating investments in a stationary market leads to an exponential growth of welfare. Welfare growing rate is dual to the market entropy. There is an obvious link between the theory of optimal investment in stock market and information theory. To explore this duality, a theory of investments can be developed.

Computation versus Communication: As we build bigger and bigger computers made by smaller and smaller components we reach both a computing limit and a communication limit. Computing is limited by communication and communication is limited by computing. These become interdependent and, therefore, all the developments in communication theory based on information theory should have a direct impact on computation theory.

2. Data Reduction

Feature subset selection is defined as a process of selecting a subset of features, d, out of the larger set of D features, which maximize the classification performance of a given procedure over all possible subsets [5]. Searching for an accurate subset of features is a difficult search problem. Search space to be explored could be very large.

The measure named *Entropy Based Discretization* is a measure commonly used in information theory, which characterizes the (im)purity of an arbitrary collection of samples, being a measure of homogeneity of samples. The entropy and information gain are functions of the probability distribution that underlie the process of communications. The entropy being a measure of uncertainty of a random variable can be used to recursively partition the values of a numeric attribute.

Given a collection S of n samples grouped in c target concepts (classes), the entropy of S relative to the classification is:

$$Entropy(S) = -\sum_{i=1}^{c} p_i \log_2(p_i) \qquad (1)$$

where p_i is the fraction of S belonging to class i.

Entropy represents the expected minimum number of bits needed to encode the class of a randomly drown sample from S. Therefore, entropy is a measure of the impurity in a collection of training samples. Using entropy an attribute effectiveness measure is defined in classifying the training data. The measure is called *information gain*, and is simply the expected reduction in entropy caused by partitioning the samples according to this attribute. More precisely, the information gain of an attribute relatively to a collection of samples S, is defined as:

$$Gain(S, A) = Entropy(S) - \sum_{v \in Values(A)} \frac{|S_v|}{|S|} Entropy(S_v) \qquad (2)$$

where *Values(A)* is the set of all possible values for attribute A, and S_v is the subset of S for which attribute A has the value v. The first parameter is just the entropy of the original collection S, and the second term is the expected value of the entropy after S is partitioned using attribute A. In other words, the information gain is therefore the expected reduction in entropy caused by knowing the value of attribute A. The information gain is

the number of bits saved when encoding the target value of an arbitrary member of S, by knowing the value of attribute A.

Using equation 2 for each feature is computed the information gain obtained if the set is split using this feature. The obtained values are between 0 and 1 being closer to 1 if the feature splits the original set in two subsets with almost the same dimensions. For selecting relevant features I use different thresholds. If the information gain obtained for a feature exceeds the threshold I will select it as being relevant, other way I will not select it.

In [4] the author justified that Information Gain failed to produce good results on an industrial text classification problem. The author says that for a large class of features scoring methods suffers a pitfall: they can be blinded by a surplus of strongly predictive features for some classes, while largely ignoring features needed to discriminate difficult classes.

3. Digital Document's Representation

Should be taken into considered that the main aim is to present the documents to interested users so that the outputs of this system should be documents that allow them to search into text as much as possible without damaging the visual information of the original documents [2].

It becomes quite clear that the documents that should be managed need to have the following characteristics:

- to contain both pictures and text
- to have the possibility of search
- to be readable by usually web clients
- to have the concept of page
- to have a small dimension
- to be supported by existing OCR systems (as output documents)

Reviewing the criteria presented in the Table 1, we have established the importance of characteristics and we take into consideration several types of existing documents including a owner document type that could be developed specifically.

The analysis concluded that the development of an owner format is not the best solution, the PDF document appears to be the best solution (see Table 1).

In addition to other documents format, the PDF format allows placing the recognized text after the original image (the text under the image) so that the user will see the pictures but will search in the text. This ensures a high quality viewing even if the recognition was a lower quality will search in the text. This ensures a high quality viewing even if the recognition was a lower quality.

The PDF documents allow also storing meta-information in them. The protections

Table 1.

Characteristics	The relevance of character	TIFF	DOC	PDF	Own format
Image + text	3	0	1	1	1
Search	4	0	0.5	1	1
Compatibility with existing customers	5	0.75	1	1	0
Page concept	2	1	0.75	1	1
Dimension	1	0.75	0.6	0.8	1
Supportability	6	1	1	1	0
Score		**12.5**	**18.1**	**20.8**	**10**

mechanisms of PDF documents are not overlooked been possible to protect a document at saving, copying, printing, copy-paste ....

The system must be thought so that all intermediate documents to be kept, if errors occurs are no longer necessary to resume the whole process (from scanning to save PDF), but only some parts of it.

Since there are many documents in the classical format, on paper, which must also be made in a digital form it is very important the OCR process (Optical Character Recognition), from our experience we recommend using one of the most successful tools in this sense – FineReader [1].

Because is used the IPA technology, FineReader allows a very good recognition of content. It can recognize text written in 177 languages. Spell check is available also in 34 languages. The recognized text can be saved in a variety of file formats, including PDF (with six options of save), HTML, Microsoft Word XML, DOC, RTF, XLS, PPT, DBF, CSV and TXT.

Portable Document Format PDF - is a file format created by Adobe Company in 1993 for exchanging the documents. The PDF standard was officially published only in 2008.

PDF combines 3 different technologies:

- a subset of PostScript programming language to describe the page, to generate graphs and the basic structure;
- a system for inclusion / removal the fonts; to allow that the fonts to be in the same document with the data;
- a structured storage system to allow these three elements and any associated content can exist together in a single file, also allow data compression.

4. Storing the Digital Documents

Multimedia means the ability to acquire, store, manipulate, combine and query information presented in more than one format, such as: text, graphics, audio, video and images. Multimedia can not be defined as a technology. It is rather a concept, which describes a number of technologies working together for the benefit of the final user. It transformed the interaction human-computer and today we have, because of it, new software products in fields like:

- access to knowledge – multimedia is probably the fastest and the cheapest way to permit access to knowledge for individuals in the manner of electronic encyclopedia
- document management – companies documents become more complex each day, containing not only text and numbers but also graphics, images, long text, etc. and multimedia can manage that diversity
- education – interactive lessons can be created for all kinds of students and disciplines
- marketing – multimedia can improve and diversify marketing activities
- real-time processes tuning and control – multimedia can be used to present in a proper way tuning and control information for real-time systems like transportation systems, patient surveillance systems, etc.

These new software products permit multimedia objects to be integrated in it and also provide a different way to visualize and interpret the labour processes. So, multimedia can extend existing applications and can change, in an innovative manner, the way we process information in different domains like economy, science, art, education and even engineering. The use of multimedia can generate benefits for all kind of users. The quality and quantity of information presented to the user is improved and also the interaction men-machine.

To organize and manage multimedia information in a suitable way we need database management systems. A multimedia database management system (MDBMS) is a preferment database management system which supports multimedia data types and can manipulate large amount of such information. A MDBMS tightly integrates three fundamental technologies like:

- database systems
- information retrieval systems
- hierarchical information storage systems.

Multimedia databases can be defined as a database system which can store, manipulate and query information presented in more than

one format such as text, audio, video, graphics, and black and white or color static images.

Multimedia databases are more and more present in today's computerized world, because they offer the possibility to easily manage different types of complex data modeled from our real each day world. Therefore, in a multimedia database we will always find new data types like:

- Image Data – these are very commonly found in multimedia databases and their applications cover simple figures, icons, medical images like X-rays, etc.;
- Video Data - these are video files and have become very important with the advent of technologies like distribution of video, etc. It is now more convenient than ever to store a home video on a personal computer.
- Audio Data - these are audio files and are being used extensively to store as well as to distribute music, sounds and speech;
- Document Data - these are the traditional text files where information is stored in the form of text. These files are still in use and have changed in terms of the capability of storage size.

Multimedia objects are different from traditional text or numerical documents in the way that multimedia objects usually require a large amount of memory and disk storage. Also, the operations applied to multimedia objects are different (e.g., displaying a picture or playing a video clip is different from displaying a text paragraph).

A multimedia database management system should be able to provide an appropriate environment for using and managing multimedia objects. Besides the traditional functions of a database management system, a multimedia database management system must be able to support the following basic functions:

- Handles image, voice, graphics and other multimedia data types
- Handles a large number of multimedia objects
- Provides a high-performance and cost-effective storage management scheme
- Provides efficient storage and retrieval of multimedia objects
- Provides efficient indexing techniques for multimedia objects
- Supports different multimedia data formats
- Supports database functions, such as insert, delete, search and update also for multimedia objects
- Provides efficient query optimizers

Multimedia objects are mostly binary large objects (BLOBs). It is common that a video clip occupies more than 100 MB of disk storage. On a video server, it is possible that thousands of video clips are stored. Due to the huge amount of storage required, a MDBMS needs a sophisticated storage management mechanism, which should also be cost-effective. The storage management scheme needs to support fundamental database operations as well.

At the same time, a MDBMS should take into consideration also the following issues:

Composition and decomposition of multimedia objects

- Operations of multimedia objects with media synchronization
- Object persistence
- Content-based multimedia information retrieval
- Concurrent access and locking mechanisms for distributed computing
- Security issues
- Consistency and referential integrity of data
- Error detection and recovery mechanisms
- Long transactions and nested transactions
- Data indexing and clustering

Unlike traditional database management systems, in a multimedia database management system data replication is not encouraged because the multimedia objects are bigger than the traditional one and this implies replication of large amount of data.

In case of relative not complicated multimedia applications, the client-server model for accessing the database can be considered appropriate. Otherwise, for complex multimedia applications it will be better to use also a specialized server (e.g. a video server)

and a multimedia database management system with a dynamic architecture.

Multimedia database management system architecture

A multimedia database usually contains three layers in its architecture:

- External layer - the user interface layer
- Conceptual layer - the object composition layer
- Internal layer - the storage layer

The tasks to be dealt with in the interface level include object browsing, query processing, and the interaction of object composition/decomposition. Object browsing allows the user to find multimedia resource entities to be reused. Through queries, either text based or visualized, the user specifies a number of conditions to the properties of resource and retrieves a list of candidate objects. Suitable objects are then reused. Multimedia resources, unlike text or numerical information, cannot be effectively located using a text based query language. Even natural language presented in a text form is hard to precisely retrieve a picture or a video with certain content. Content-based information retrieval research focus on the mechanism that allows the user to effectively find reusable multimedia objects, including pictures, sound, video, and other forms. After the successful retrieval, the database interface should help the user to compose/decompose multimedia documents.

The second layer work in conjunction with the interface layer to manage objects. Typically, object composition requires a number of links, such as association links, similarity links, and inheritance links of an object-oriented system to specify different relations among objects. These links are specified either via the database graphical user interface, or via a number of application program interface (API) functions.

The last layer, the storage management layer, includes two performance related issues: clustering and indexing. Clustering means to organize multimedia information physically on a hard disk (or an optical storage) such that, when retrieved, the system is able to access the large binary data efficiently [7]. Usually, the performance of retrieval needs to guarantee some sort of Quality of Service and to achieve multimedia synchronization. Indexing mean that a fast locating mechanism is essential to find the physical address of a multimedia object. Sometimes, the scheme involves a complex data or file structure.

5. Case Study: SCRIBe

As part of the SCRBEe project [6] "Information system for processing and visualization of old books inventory" in the university library of Sibiu a number of 20 old books was digitized. The project was meant to capitalize on existent books and make them available in the new format to readers, by scanning and archiving them in electronic form.

The main objective of SCRIBe [6] was to realize an information system to allow the beneficiaries (citizens, Romanian and foreign researchers, libraries and museums staff) the access to old books inventory, because normally the access is restricted both because of the sparseness of the copies and because of the need to protect copies with a high degree of degradation. The interest for those books is given by their age, their rarity and by scientific reasons related to history, linguistics and theology (mainly orthodoxies).

The SCRIBe objectives were:

- Realization of an experimental information system for acquisition, compression and management of the documents' images;
- Building a WEB site where users can search in the virtual library of old books and express their interest in reading some of them;
- Realization of an experimental OCR system which will be adaptive to the different types of writings in that documents;
- Realization of a system that will allow users to visualize different areas from images at different details levels;
- Designing and implementation of an experimental system that will allow users to apply different methods for image enhancement of some areas and to express their satisfaction about them;

Technological structure and electronic resources used in digitization process:

- Technological structure:
  - a Pentium 4 PC – 2,4 GHz, 512Mb RAM, 80Gb HDD,
  - a HP ScanJet 4370 with scanning software,
  - books with Latin and Cyrillic characters, in Romanian, German, French, Russian languages,
- Electronic resources:
  - saving in Acrobat Reader pdf, jpg images, and some books in Word format, with character recognition,
  - introducing in a database available on web, in xml format,
  - books can be searched and visualized.

As part of that project a number of 20 books were scanned (presented in annex 1). The first 20 pages from each book were automatically transformed in xml format in order to allow indexing and search of information from those books. In choosing that number of pages copyrights reasons were also taken into account. The books were also saved in jpg and pdf formats, for their later use in digitization projects.

The criteria that formed the basis for selecting publications in order to realize digital collection of the library were the degradation risk, value and rarity of publications from the library inventory.

6. Conclusions

The experience gained in SCRIBe realization highlighted the difficulties of storage, retrieval and shared access to digitized documents. The use of modern methods for data reduction based on information theory is now the most rigorous method for realization of digital archives. The future challenges have also been highlighted namely the integration of the various formats present in multimedia documents. What a wonderful thing will be to realize images retrieval in a video archive based on a combination of search words.

Acknowledgements

This work was partially supported by the Romanian National Council of Academic Research (CNCSIS) through the grant CNCSIS no. 12099/2008-2011.

REFERENCES

1. http://finereader.abbyy.com/
2. BANCIU, D., **e-Romania - A Citizens' Gateway towards Public Information**, Journal of Studies In Informatics and Control, Vol. 18, No. 3, 2009
3. COVER, T. M., T. A. JOY, **Elements of Information Theory**, John Wiley & Sons Interscience Publication, 1991.
4. FORMAN, G., **A Pitfall and Solution in Multi-Class Feature Selection for Text Classification**, Proceedings of the 21st International Conference on Machine Learning, Banff, Canada, 2004.
5. GUERRA-SALCEDO, C., S. CHEN, D. WHITLEY, S. SMITH, **Fast and Accurate Feature Selection Using Hybrid Genetic Strategies**, CEC00, Proc. of the Congress on Evolutionary Computation, CEC00, July 2000.
6. http://alice.ulbsibiu.ro:8080/scribe/
7. SIFAOUI, A., A. ABDELKRIM, S. ALOUANE, M. BENREJEB, **On New RBF Neural Network Construction Algorithm for Classification**, Journal of Studies In Informatics and Control, Vol. 18, No. 2, 2009

Annex 1 – List of scanned books in the SCRIBe project

No.	AUTORS	TITLE	PLACE	PUBLISHER	YEAR
1.	Ioane V. Rusu	Compendiu de Istoria Transilvaniei	Sibiiu	Tiparitu la S. Filtsch	1864
2.	Tim. Cipariu - Canonicu Gr. Catolicu	Elemente de limb'a Romana dupa dialecte si monumente vechi	Blasiu	cu tipariulu sem. Diecesanu MDCCCLIV	1854
3.	Alexandru Philippide	Introducere in Istoria limbei si Literaturei romane	Iasi	Editura librariei Fratii Saraga	1888
4	Dupa Deregintele Preparandialu Ignatiu BARANY	CRESCEREA POPORALA - Manualu Pedagogicu-Didacticu	Oradea Mare	Tipariulu lui Eugeniu Hollosy	1879
5	WILHELM BRAUNE	ALTHOCHDEUTSCHE GRAMMATIK	HALLE	MAX NIEMAYER	1891
6	Karl Prochasta	Lessings - stliche lyrische, epische und dramatische Werke und seine vorzen Prosaschriften	Leipzig	Leipzig und Teschen	1806
7	Heinrich Laube	Lessing s Werke	Wien, Leipzig, Prag.	Verlag von Sigmund Bensinger	1895
8	-	Die Geschichten des Herodotos	Leipzig	Druck un Verlag von Philipp Reclam jun.	1885
9	Johann H. Vok	Homers Werke	Stuttgart	Verlag der J Gottaschen	1869
10	-	Codicele civil	Bucuresti	Tipo-Litografia Ed. Wiegand & C. Savoiu, Covaci	1894
11.	Manolaki D.	Istoria Moldovei pe timii de 500 ani pana in zilele noastre	Iasi	Tipografia Institutul Albinei	1857
12	-	- Motii - Rascoala romanilor ardeleni 1784 - 1785 sub capetenia lui Horia - Curcanii - Luarea Rahovei de ostile romanesti la Noemvrie 1877	Bucuresti	Editura Librariei Leon Alcalay, No. 37 - Calea Victoriei	1877
13	Studiu comparativ de Lazar Sainenu	Basmele Romane in comparatiune cu legendele antice clasice si in legatura cu basmele poporeloru invecinate si ale tuturoru poporeloru romanice	Bucuresci	Lito-Tipografia Carol Gobl	1895
14	Michael J. Ackner	Die Romischen Inschriften in Dacien	Wien	Verlag Tendler &Co.	1865
15		Histoire des Roumains et de leur civilisation		Cvltvra Nationala Bucarest	1922
16	A.D. Xenopol	Istoria romanilor din Dacia Traiana	Bucuresci	Ed. Librariei Scoalelor	1914
17		Ce sint si ce vor sasii din Ardeal Expunere din izvor competent	Bucuresti	Tipografia "Cultura Neamului Romanesc"	1919
18.	Gheorghie Sincai	Chronica Romanilor si a mai multor nemuri	Bucuresci	Tipografia Academiei Romane (Laboratorii Romani)	1886
19	-	Istoria Literaturilor Romanice in Desvoltarea si Legaturile lor Vol I-iu Evul Mediu	Bucuresti	Tipografia "Cultura Neamului Romanesc"	1920
20	-	Graiul Nostru	Bucuresti	Atelierele Grafice Socec and Co., Societate anonima	1908