Classification and automatic indexing in a persistent object environment

Robert Godin and Brigitte Kerhervé
Université du Québec à Montréal
principal investigators

James Turner
Université de Montréal
co-investigator

Goals and objectives

The goal of this project is to study classification and automatic indexing for multimedia data in the context of digital libraries. The general objectives are: (1) to propose methods for digital media classification and indexing and (2) to implement them effectively by using a persistent environment provided by object-oriented database systems.

Background and rationale

"Digital libraries" is one of those confusing buzzwords in current use. The confusion arises from its use in various contexts, each with its own meaning. "To some it simply suggests computerization of traditional libraries... [to others] a distributed text-based information system..., a collection of distributed information services..., a distributed space of interlinked information..., or a networked multimedia information system... [1]. In this paper it mostly refers to distributed multimedia information systems.

Libraries of the future will integrate different types of data such as texts, graphics, images, video and audio in multimedia information systems [2]. Typically such systems must manage a very large amount of data, coming from different sources and generally stored in distributed databases [3]. Potential users of such systems require effective access to pertinent information using appropriate retrieval

ISSN: 2324-9773

Godin, R., Kerherve, B., Turner, J., Turner, J., & Turner, J. (1995). Classification and automatic indexing in a persistent object environment. 6th ASIS SIG/CR Classification Research Workshop, 53-64. doi:10.7152/acro.v6i1.12661

strategies. Due to the particularities of multimedia data, there is a need to investigate various issues in order to achieve these goals. Images, audio, and video data in digital form are voluminous, and they lack the alphabet which makes text searching possible.

The first issue to be resolved is to define abstractions of raw data through the description of metadata, i.e. data about data [4]. Such abstractions can be textual descriptions of the content as well as the association of visual elements to digital media. The second issue deals with the organization of these descriptions and their combination to offer various search strategies. The third issue concentrates on the effective storage and access of this metadata in a persistent environment, in order to permit sharing and re-using existing descriptions [5].

Defining abstractions of raw data appears in conjunction with the need for indexing large text databases for purposes of storage and retrieval. In the 1960s, when the first large text databases were being built, research into automatic indexing was begun. Techniques reflected the computer technology available at the time, and the KWIC (key word in context), KWOC (key word out of context), and KWAC (key word and context) type indexes soon became a common feature of specific types of published materials. This style of text index results from a simple sort of all the words in a text, after removal via a stop list of words without significance for indexing, such as articles, pronouns, and conjunctions. The next generation of automatic indexing research built upon these early techniques by adding more sophisticated interventions, such as counting occurrences of words, adding term weighting, and clustering techniques.

Multimedia data present a new set of problems in indexing because of the simple but unavoidable fact that still image, sound, and video data do not contain text which can be filtered through algorithms in order to derive meaning and thus apply automatic indexing and classification techniques [6]. Yet for purposes of storage and retrieval, words are always associated with still and moving pictures and with sound material. One of the important reasons for this is that users looking for such material in information systems almost invariably arrive with a verbal request. Some experimental research into using visual elements to retrieve pictures has been undertaken, e.g. [7, 8, 9], and has met with some success. However, the general conclusion that can be reached from this is that while types of access other than strictly verbal access to multimedia information objects may be desirable in some circumstances, it is neither possible nor desirable to eliminate textual indexing of non-text information objects [10, 11].

Indexing is the preliminary step to allow effective access to multimedia objects. Whether it is generated manually or automatically, the index is used in direct querying as a retrieval mechanism. Such a mechanism is generally not sufficient, since users should be provided with a variety of interfaces for exploring the multimedia database. Thus it is useful to combine indexing with classification techniques in order to provide the users with other search strategies. Our work on using automatic conceptual clustering of documents for retrieval purposes is an approach to this problem [12]. The conceptual hierarchy generated is used as a navigational space where browsing and direct querying can be integrated in a complementary and coherent fashion. This is particularly attractive for non-expert users having little experience with a particular database or representation language.

As pointed out by many researchers, e.g. [13, 14, 15, 16], an information retrieval system should provide browsing mechanisms for users who do not know precisely what they want or how to get it. Browsing is based on freely exploring a structure such as a tree or a graph in order to find useful items. Browsing in some form of data space has long been recognized as an attractive alternative for information retrieval particularly for casual users and exploration of new domains [17]. In addition, it is widely recognized as an important strategy in retrieving pictures, e.g. [18, 19, 20].

Our initial work used the Galois lattice of an indexing relation as a medium to support browsing [21]. One important advantage over other browsing paradigms such as navigation in an enumerated classification structure or a hypertext structure is that the browsing space is not manually built but automatically generated. Another advantage is the natural integration of direct querying and browsing modes of interaction because they are based on the same retrieval space. Small scale experiments with real users have shown the potential of the method compared to more traditional methods such as Boolean querying and navigating in manual classification schemes [22, 23] As a generalization of the previous work, the knowledge space structure generated from conceptual graph representations has also been proposed as a browsing space for retrieval in the context of image databases [24]. A recent effort is directed towards combining automatic indexing methods

ISSN: 2324-9773

based on high-level linguistic analysis of texts and conceptual clustering using the indices derived [25]. All these issues must be addressed in the context of database systems which allow sharing and re-using existing objects.

Several efforts have been recently dedicated to developing distributed multimedia systems which integrate services for multimedia object creation, storage, access, transfer and presentation [26, 27]. Such systems integrate various components, and database systems play a major role among them in providing a reliable and persistent environment for multimedia obejct storage and access. Database systems allow sharing and reusing existing objects, functions that are essential for digital libraries [28].

Information stored in a multimedia database falls into two categories: multimedia objects accessed by the users, and metadata used by the system to search, access, transfer or present multimedia objects effectively [29]. In the database community, research efforts concerning multimedia documents have essentially focussed on the first category and have led to propositions for modelling multimedia objects [30], for query languages [31] and for multimedia object storage [32]. Klas [33] recently identified the importance of definition, organization, storage and access to metadata for digital media within the framework of distributed multimedia systems. Some propositions have been made for managing multimedia documents [34], images [35] or videos [36]. Several issues need to be adressed concerning the metadata for image classification and indexing in the context of digital libraries. These issues include organization of metadata, efficient use of the access primitives provided by the database system, and access to remote databases and to their metadata.

Research Plan

This project focusses on metadata for content description of digital media through classification and automatic indexing. The first stages of the project will concentrate on classification and indexing of still images and text. This will subsequently be extended to video as well as to multimedia documents in general. Classification and indexing will be processed in a persistent object environment using object oriented database technology. Three issues will be addressed in this

ISSN: 2324-9773

project: identification and representation of concepts for indexing, metadata for classification and indexing, and management of the persistent object environment

Identification and representation of concepts for indexing

Multimedia information systems by nature include large volumes of data. There is no way of getting around the problem of organizing this data in highly systematic ways if retrieval from such systems is to be efficient. Thus the need arises for appropriate indexing techniques which must be successful yet automated as far as possible in order to ensure that they are cost-effective. Essentially, effective ways for relating words to nontextual information objects is what is required in order for users of such systems to be satisfied.

Criteria need to be developed for deciding what constitutes an object which is significant enough in images or texts to justify indexing it. In addition, the problem of naming information objects needs to be addressed. One direction is automatic indexing of texts using a combination of natural language processing and traditional statistical methods. Another direction is linking textual representations of indexing concepts to pictorial representations of the concepts in a classification of the world showing whole-part relationships, in order to allow users to use words and pictures together to retrieve visual information objects.

It has been shown that users of visual information systems tend to simply name persons, objects, or events about which they are seeking visual information [37]. In order to provide the links between words and pictures, a controlled, structured vocabulary needs to be available to users. In addition, visual information available alongside the textual information would help provide enough context to reassure users that the system understands their request.

One strategy for linking indexing terms in the form of words and phrases to images in a multimedia information system would be to use a visual dictionary in conjunction with the system [38]. The dictionary is produced from computer files which include high-quality graphic renderings of objects with textual labels in the various languages available. One objective of this part of our research project is to test the effectiveness of the dictionary as a front end to the information system. A

ISSN: 2324-9773

Godin, R., Kerherve, B., Turner, J., Turner, J., & Turner, J. (1995). Classification and automatic indexing in a persistent object environment. 6th ASIS SIG/CR Classification Research Workshop, 53-64. doi:10.7152/acro.v6i1.12661

second objective is to test the effectiveness of the whole-part text component as a controlled vocabulary for the system. A third objective is to determine whether the classificatory structure superimposed on the pictures included in the dictionary can be applied successfully to a multimedia information system with either general content or specialised content.

Metadata for classification and indexing

In this issue, we will study metadata to support querying and browsing of texts and images represented by concepts, as well as taxonomic relationships among the concepts. The metadata are the result of automatic conceptual clustering of the texts and images using representation of the basic content. We intend to explore strategies for conceptual clustering for text elements and visual elements. In particular, we plan to explore conceptual clustering based on concept lattices and variant structures using the content representation addressed above.

Management of the persistent object environment

Metadata for classification and indexing will be stored in an object oriented database system. This issue deals with the representation, storage and access of metadata and with the efficient interaction with the database system in a distributed environment.

Metadata for image indexing and classification can include textual as well as visual representations. These representations are shared among all users and must be stored and managed by the database system. One of the early activities in this research is designing the database schema for metadata. This schema needs to integrate in a uniform way the various kinds of metadata, such as that for the representation of types of media, for the structure of multimedia objects, for the description and classification of the content, for storage, and for different versions of documents. We essentially focus on metadata for content description and classification, but intend to consider the other categories also, in order to provide a general framework for metadata for distributed multimedia databases.

Metadata will be used as the key link to pertinent multimedia objects and thus

ISSN: 2324-9773

need to be extremely well organised in order to be effective. We will need to propose a detailed design of the metadatabase including the physical schema as well as the behavioral schema. We propose to use object-oriented database technology and to implement the metadatabase in an existing object-oriented database system.

Metadata are used during the querying process in order to reduce the search space to explore. Since we plan to include both direct querying and browsing, we need to study and propose approaches to manage the interactions between the querying process and the database system. In particular, we intend to explore strategies to reduce the calling of primitives within the database system.

Context of the larger research proposal

All these issues need to be considered within the framework of the larger research proposal, in which a distributed environment is used, hence the need to investigate effective access to distributed metadatabases. The larger proposal includes five projects. Three of these focus on ways of building collections and searching through them. These cover text and document libraries, image and graphic libraries, and music and acoustic signal libraries. A fourth project will investigate AI techniques for visualization and interaction based on the agent paradigm with the purpose of helping in building more advanced systems for exploring and using digital library data. The fifth project is concerned with system architecture and integration of the components developed by the other projects.

Milestones in the image and graphics project

Year 1: Identifying textual and visual concepts in a specific image collection; installing the visual dictionary and making links with pictures in the system; concept clustering for visual and textual elements; designing the metadatabase schema; developing the first image classification tool prototype

Year 2: devising benchmark searches to run against the system; devising ways to measure the success or failure of retrieval; combining indexing and classification structures in the search process; optimising primitive calls within the database

ISSN: 2324-9773

system; extending the metadatabase to the distributed environment; exploring the applicability of image classification techniques to videos.

References

(1) Edward A.Fox et al., [Introduction to the theme section on digital libraries], *Communications of the ACM* 38, no. 4 (April 1995) 24–28.

(2) E.A. Fox, "Digital libraries", *IEEE Computer,* 26, no. 11 (November 1993) 79-81.

(3) P.B. Berra et. al., "Architecture for distributed multimedia database systems", *Computer Communications,* 13, no. 4 (May 1990).

(4) W.Klas, "Metadata for digital media: introduction to the special issue", *ACM Sigmod Record,* 23, no 4 (December 1994) 19–20.

(5) W. Grosky et al., "Using metadata for the intelligent browsing of structured media objects", *ACM Sigmod Record,* 23, no 4 (December 1994) 49–56.

(6) James Ian Marc Turner, "Determining the subject content of still and moving image documents for storage and retrieval: an experimental investigation" (PhD thesis, University of Toronto, 1994).

(7) M. E. Rorvig, "The substitutability of images for textual descriptions of archival materials in an MS–DOS environment" *The application of microcomputers in information, documentation, and libraries,* ed. K. Lehman and H. Strohl-Goebel (Amsterdam: North Holland Press, 1987) 407–15.

(8) Gary A. Seloff, "Automated access to the NASA-JSC Image Archives" *Library Trends,* 38, no. 4 (spring 1990) 682–96.

(9) Howard Besser, "Visual access to visual images: the UC Berkeley Image Database Project", *Library Trends,* 38, no. 4 (spring 1990) 787–98.

ISSN: 2324-9773

(10) James Turner, "Representing and accessing information in the stockshot database at the National Film Board of Canada", *Canadian Journal of Information Science,* 15, no. 4 (December 1990) 1–22.

(11) James Ian Marc Turner, "Determining the subject content..." .

(12) R. Godin et al., "Experimental comparison of three methods for personal information retrieval", in *Second Annual Symposium on Document Amalysis and Information Retrieval, Las Vegas, Nevada* (1993).

(13) R.H. Fowler, W.A. Fowler, and B.A. Wilson, "Integrating query, thesaurus and documents through a common visual representation", in *14th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Chicago* (1991).

(14) T.K. Landauer, S.T. Dumais, L.M. Gomez, and G.W. Furnas, G. W., "Human factors in data access", *The Bell System Technical Journal,* 61 (1982), 2487–2509.

(15) G. Marchionini, and B. Shneiderman, "Binding facts vs. browsing knowledge in hypertext systems", *IEEE Computer,* 21, no. 1 (1988) 70–80.

(16) R.H. Thompson, and B. Croft, "Support for browsing in an intelligent text retrieval system", *International Journal of Man-Machine Studies,* 30 (1989) 639–668.

(17) G. Marchionini, and B. Shneiderman, "Binding facts vs. browsing knowledge..."

(18) Brian O'Connor, "Moving image-based serial publications", *Serials Review,* 12, nos. 2-3 (summer/fall 1986) 19–24.

(19) Estelle Jussim, "The research uses of visual information", *Library Trends,* 25, no.4 (April 1977) 763-778.

(20) Nancy Shelby Schuller, "Classification system for images of church interiors", *Visual Resources Association Bulletin,* 20, no. 2 (summer 1993) 23–24. Schuller 1993, 23).

ISSN: 2324-9773

Godin, R., Kerherve, B., Turner, J., Turner, J., & Turner, J. (1995). Classification and automatic indexing in a persistent object environment. 6th ASIS SIG/CR Classification Research Workshop, 53-64. doi:10.7152/acro.v6i1.12661

(21) R. Godin, J. Gecsei, and C. Pichet, "Design of a browsing interface for information retrieval", in *12th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Cambridge, MA (1989).*

(22) R. Godin et al., "Experimental comparison of three methods... "

(23) R. Godin, R. Missaoui, and A. April, A., "Experimental comparison of navigation in a Galois lattice with conventional information retrieval methods", *International Journal of Man-Machine Studies,* 38 (1993) 747–767.

(24) G. Mineau, and R.Godin, "Automatic knowledge structuring for browsing retrieval", in *International Conference on Information and Knowledge Management, Baltimore* (1992) 273–281.

(25) R. Godin et al., "Applying concept formation methods to software reuse", *International Journal of Knowledge Engineering and Software Engineering.* [in press 1995].

(26) P.B. Berra et al., [Guest Editor, Introduction to multimedia information systems], *IEEE Transactions on Knowledge and Data Engineering,* 5, no. 4 (August 1993).

(27) B. Kerhervé et al., "Functional requirements for a generic distributed multimedia presentational application", in *Third International Conference on Computer Communications and Networks, San Francisco (1994).*

(28) E.A. Fox, "Digital libraries".

(29) B. Kerhervé, B., and L. Ouédraogo, "Méta-informations pour les bases de données multimédia réparties", [in press 1995].

(30) C. Meghini, F. Rabitti, and C. Thanos, "Conceptual modeling of multimedia documents", *IEEE Computer* (October 1991).

ISSN: 2324-9773

(31) F. Golshani, and N. Dimitrova, "Design and specification of EVA: a language for multimedia database systems", in *3rd International Conference on Database and Expert Systems, Valencia, Spain.* Springer Verlag (1992).

(32) T.V. Rangan, and H.M. Vin, "Efficient storage techniques for digital continuous multimedia", *IEEE Transactions on Knowledge and Data Engineering,* 5, no 4 (August 1993).

(33) W.Klas, "Metadata for digital media...".

(34) K. Bohms, and T. Rakow, "Metadata for multimedia documents", *ACM Sigmod Record,* 23, no 4 (December 1994) 21–26.

(35) Y. Kiyoki, T. Kitagawa, and T. Hayama, "A meta-database system for semantic image search by a mathematical model of meaning", *ACM Sigmod Record,* 23, no 4 (December 1994) 34–41.

(36) R. Jain, and A. Hampapur, "Metadata in video databases", *ACM Sigmod Record,* 23, no 4 (December 1994) 27–33.

(37) James Ian Marc Turner, "Determining the subject content..." .

(38) Jean-Claude Corbeil, and Ariane Archambault. *Le visuel multilingue: dictionnaire thématique.* Montréal: Éditions Québec-Amérique (1994).

ISSN: 2324-9773

ISSN: 2324-9773