BULGARIAN ACADEMY OF SCIENCES
INSTITUTE OF MATHEMATICS

CONFERENCE ON INTELLIGENT MANAGEMENT SYSTEMS

1

September 26 — October 2, 1989
Varna, Resort Druzhba, Bulgaria

Managing Uncertainty in a Multimedia Document System

*P. Conti**, F. Rabitti*, P. Savino**, P. Stanchev****

* IEI-CNR, Pisa
** Olivetti DOR, Pisa
*** Bulgarian Academy of Sciences, Sofia

ABSTRACT

This paper describes how to take into account the uncertainty factor in the retrieval of multimedia documents. Uncertainty is mainly introduced in evaluating how images and text components in documents are relevant to the user's queries. A suitable management of uncertainty can significantly increase the effectiveness in the retrieval process.

Uncertainty management is discussed in the context of project MULTOS, a prototype system for the storage and retrieval of multimedia documents.

1. Introduction.

A key problem in all application environments where documents have to be managed is the increasing amount of information that is being generated with the wide use of documents in electronic form. This amount of data is likely to increase with the introduction of tools allowing an effective management of multimedia documents.

A multimedia document can be defined as a collection of components which may contain different information in form of text, formatted attribute, image and voice. The components can be mixed and interrelated, and they may have an internal structure. As a result, these documents have complex structures which tend to differ from one document to another.

Editors for multimedia document creation and formatting[8] are actually available, while multimedia document exchange is allowed by the presence of tools for multimedia mails and standards for document transmission[11].

However the increasing number of documents in electronic form makes their efficient storage and their efficient and effective retrieval a central issue.

These two central problems are addressed in project MULTOS. MULTOS (MULtimedia Office Server) is an ESPRIT Project in the area Office Systems. It supports basic filing operations, such as creation, modification, and deletion of multimedia documents, and the ability to process queries on documents. Documents are stored in data bases and may be shared by several users while facilities like

This work has been funded in part by the Commission of the European Communities as ESPRIT-Project No. 28. Partner in this Project are Ing. C. Olivetti S.p.A. (I), Cretan Research Center (GR), IEI-Consiglio Nazionale delle Ricerche (I), TA Triumph-Adler AG (D), Battelle Institut e.V. (D), Mnemonica Computer Services Ltd. (GR) Estudios y Realizaciones en Informatica Aplicada, S. A. (E)

authorization, version, and concurrency control are supported.
An important system feature is the possibility to store very large amounts of data; these facility is supported integrating Optical Disk Media into the system.
Multos is based on a client/server architecture. Three different types of document servers are supported, *current server, dynamic server and archive server.* They allow filing and retrieval of multimedia documents based on document collections, document types, document types attributes as well as on document content (text, im ge).

The retrieval approach adopted in Multos provides a fast retrieval and makes possible, during the formulation of the query, to express many of the user needs.
However the approach has the disadvantages typical of the methods based on exact matching; if a small error in matching occurs, the document is not retrieved; furthermore no evaluation of the different relevance of the retrieved documents is provided to the user.
Moreover it is not possible to express the uncertainty that the user has on part of the query he is formulating. Many relevant documents are disregarded and many non relevant ones are retrieved.

These topics have been addressed in the research of Information Retrieval: document retrieval is viewed as a process of plausible inference [14]. A document D is retrieved if it logically implies the query Q. For example if Q = "*signature based retrieval techniques*" and D contains the phrase "*text retrieval is based on Superimposed Coding techniques*" then Q can be inferred from D. However the documents rarely imply the exact query and there is always an uncertainty associated with the implication.
The formalization of this approach can be done in various ways (e.g. using a probabilistic model[13], vector space model[16], etc.). All these methods allow for the retrieval of a list of documents ranked in order of the certainty of the inference.

In this paper, we first describe (in Sec.2) the main features of prototype MULTOS, then we discuss (in Sec.3) the techniques for multimedia document retrieval, based on document content specification. The shortcomings of the actual MULTOS approach, caused by the lack of support for uncertainty management, are then discussed. Then, we present the techniques to be introduced in MULTOS to support partial match in the access by content for text components (in Sec.4) and for images (in Sec.5). In Sec.6 we discuss how the previous partial match techniques, for image and text, can be integrated in managing uncertainty in the proposed extension of MULTOS. Final remarks are given in Sec.7.

2. MULTOS Filing and Retrieval Capabilities

In order to support an effective and powerful document retrieval the system must be able to answer queries at a high level of abstraction, where conditions on different documents components, such as free text, formatted attributes and images, are intermixed. An example of a typical query in an office environment is:

> *Find all the offer letters about PCs with a price of approximately $ 5,000 produced by Olivetti. These PCs are described as products having good ergonomics. I am rather sure about the producer and the price but not about the description. Moreover, the offer letter should contain a picture of the PC, complete of screen and keyboard, and probably with two floppy drives.*

This typical query can be expressed if we provide a model that allows to represent high level concepts (e.g. price) present in the documents contained in the database, to group documents into classes of documents having similar content and structures (e.g. offer letter) and that allows to express free text conditions (e.g. good ergonomics).

The MULTOS document model has been developed with the purpose to support operations such as editing, presentation, transmission and retrieval of multimedia documents.
This is obtained allowing the description of the document logical, layout and conceptual structures. The **logical structure** determines how the document logical components (e.g. title, introduction, chapter, section, etc.) are arranged. The **layout structure** defines the layout of document content at output. It contains components such as page, frame, etc. The **conceptual structure** allows a semantic oriented description of document content as opposed to a syntax oriented description provided by logical and layout structures.

The logical and layout structures are defined according to the ODA[11] document representation. The conceptual structure has been added to provide support for document retrieval by content.

Documents having similar conceptual structures can be grouped into classes (conceptual types).
In order to handle types in an effective manner, types are maintained in a hierarchy of generalization, where a subtype inherits from its supertype the conceptual structure that then will be further refined.

For *document retrieval* the conceptual types play the role of a data base schema that enables the maintenance of efficient access structures and that is the basis for formulating queries on an abstract level.

Since an open system must be capable of handling documents from arbitrary sources, including documents without a conceptual structure, a stepwise construction of the conceptual structure by an editor during the creation of the document is not sufficient. Therefore the MULTOS system provides a knowledge based classification subsystem (In our sense "classification" not only comprises the association of a document to a class but also the generation of the conceptual structure from raw documents) that analyses the document's content and automatically constructs a conceptual structure based on a given set of type definitions[2, 7].

Document retrieval is based on a boolean query language that allows to specify restrictions on document collections, document types, document attributes and free text.

The user query previously presented becomes:

```
FIND DOCUMENTS SCOPE = PC-documentation TYPE = offer-letter WHERE
product-price = 5000 AND company-name = Olivetti
AND TEXT CONTAINS good ergonomics
AND WITH IMAGE;
```

It can be observed that the user query cannot be exactly translated in the actual MULTOS query language. In particular the uncertainty the user has about some attribute values cannot be expressed. Also the condition on the image content cannot be expressed, only the presence of one image in the document may be requested.

3. Techniques for Multimedia Document Retrieval.

From the previous discussion derives that an efficient retrieval can be obtained using the type hierarchy and the relations between instances and types for narrowing the number of instances that must be checked for matching. However it is important for the user to have the possibility to specify part of the content of the required documents. Document content is represented by raster images, voice and text.

The problem of *image retrieval* by content cannot be approached trying to perform picture recognition that involves very expensive pattern recognition routines. Furthermore it is very difficult for the user to specify precisely the content of the picture he/she wants.

Voice retrieval through speech recognition presents similar problems. Currently only speaker dependent, discrete speech voice recognition devices with a limited vocabulary of words exist in the market.

These difficulties can be somewhat bypassed by retrieving audio and image data through associated text or attribute information and through some of their broad characteristics (e.g. position of an image, name of a speaker, etc.).

However, true content addressability for these types of data is a desirable goal. This is especially true for images as we expect the document filing system to handle much more image data in comparison to voice.

Many methods are being tried to solve the problem of Multimedia document retrieval. At a general level, they form two classes: exact match techniques and partial match techniques.

The *exact match* retrieval techniques provide a basic token matching capability in that only the documents that exactly match with the specified query can be retrieved.

The *partial match* retrieval techniques allow to retrieve the documents that match only partially with the query.

The first group of techniques are widely used in the existing retrieval systems; they mainly address the problem of efficiency, which is crucial because of the dimension of the document base to be handled. Moreover these techniques are all very well experimented and their usefulness has been proven in real user environments (where seems that important aspects of user' s queries are well represented by Boolean statements).

However the exact match retrieval techniques have some disadvantages[1]:

a) documents whose representation matches the query only partially are missed

b) the documents cannot be ranked in relevance order

c) the relative importance of concepts either within the query or within the text cannot be taken into account.

The partial match techniques are more powerful of the exact match techniques for what concerns the effectiveness of query results, but they have never been experimented in real environments and their use could be problematic in very dynamic environments like the office one, because the insertion/deletion of new documents has an high cost. A recent review of the partial match techniques can be found in[1].

Recently an interesting method that allows to express the uncertainty that the user has on the restrictions he is posing[6, 10] has been presented. The query is formulated in a non-boolean way and the results are returned in decreasing relevance order, based on the certainty the user has on the restrictions he expressed.

The actual MULTOS system adopts an approach based on exact match for formatted attributes and text components that, as it will be described in the next chapter, can be extended to support partial match and user uncertainty in query formulation.

4. Retrieval of Text Components.

A comparison of the different methods that can be used for text retrieval[12] for what concern their adaptability to an office environment, where frequent insertion are performed, and where a good response time is required for text queries, gave the result that the signature method with superimposed coding (SC)[3] is the most suitable.

This method of signature extraction allows simple management of insertions, has a small space overhead (8-12%), and has a good response time for the query (about 20 sec. for a data base of 90Mbytes).

Moreover it provides for automatic elimination of duplicate words within a block of text and gives the possibility to improve the efficiency performing in parallel different queries accessing only one time the signature file. This method can be extended to provide partial match for the document textual components and to rank the retrieved documents in decreasing relevance order [5].

Ranking strategies associate a score which depends on an evaluation of the relevance of that document for the query with each document in the collection. The probabilistic retrieval model provides a formal basis for the ranking strategies. Each document D is represented by a set of index terms with associated weights (D = $(t_1,...,t_n)$ with t_i= 0 or 1 indicates if term i is present or absent in the document) and each query is represented by Q = $(qt_1,...,qt_n)$, where qt_i indicates whether or not the term i is present in the query. The probabilistic model allows for the calculation of optimal ranking functions [15]. Considering the textual component of the document, these can be estimated by calculating some term frequency parameters (e.g. term frequencies in the document collection and within document term frequencies) in the document collection. Basically the following ranking functions, given in increasing complexity order and in increasing relevance order, can be obtained:

a) $\sum_{i|qt_i=1}^{n} w_i t_i$

where $w_i = \log \dfrac{nd}{f(t_i)}$

nd = number of documents stored

and $f(t_i)$ = number of documents containing term t_i.

b) if within-document term frequency is also included, better performance can be obtained

$\sum_{i|qt_i=1}^{n} s_{ij} w_i t_i$

where s_{ij} = frequency of term i in document j.

c) further improvement of retrieval effectiveness can be obtained if a correction factor is considered for documents containing dependent groups of terms.

Adopting a simple variation of the signature algorithm, it is possible to obtain a performance comparable to that obtainable in (c) [5]. Two different ranking algorithms will be considered; one for a sequential organization of the signature file and the other for a bit-sliced organization. Here we present the algorithm followed when a sequential organization of the Signature file is used.

Consider a query $Q = (qt_1,...,qt_k)$ with k terms. If a sequential organization of the Signature file is used, the following procedure has to be followed:

- For each qt_i $i = 1,...,k$ generate the signature for the stem corresponding to qt_i, obtaining $S_Q = (s_{qt_1},...,s_{qt_k})$.
- Scan the signature file. For each text block signature compare it with s_{qt_j} $j = 1,...,k$ and maintain the result in temporary storage.
- When a block signature for a new document is encountered, the score for the previous document is calculated through

$$\sum_{i=1}^{k} \frac{bc_i}{b} w_i$$

where b = number of block signature of the document

and bc_i = number of blocks in the document containing qt_i.

$\dfrac{bc_i}{b}$ gives an estimate of the within-document term frequency. This estimate is more precise for long documents than for short documents.

- The document score can be adjusted according to the presence of document groups of terms, such as those specified in a Boolean query. For each dependent group of terms a signature is generated and compared with the block signatures. A score is added to the documents containing the group of terms.

Relative to an exact match strategy using the sequential signature file, the ranking strategy requires almost the same number of I/O operations, the only additional access required being those to the term posting file. The processing time is slightly increased, because an exact match strategy uses only one signature to represent each conjunct in a boolean query, whereas the ranking strategy uses a signature for each query term, and calculates the document scores. However the difference is not significative since the response time depends mainly from the time required to scan the signature file.

In [5] a more detailed description of the algorithm is presented together with a comparison of its performance with classical term-based document representations.

5. Retrieval of Image Components

The main conceptual problem in dealing with images derives from the difficulty to exactly define and interpret the content of images. Images can be very rich in semantics, but are subject to different interpretations according to the human perspective or the application domain. On one hand, it is difficult to recognize the objects (with the associated interpretation) contained in an image, on the other hand is difficult to determine and represent the mutual relationships among these objects, since they form structures which vary greatly from image to image.

An approach could be to apply DBMS or IRS techniques. However, with respect to DBMS's, it is difficult to recognize regular structures of objects contained in images, and then organize image instances into a limited number of types, to which the interpretation is associated. This is the approach required by the strictly typed data models adopted in database systems. In IRS, instead, a free formatting of text is allowed, usually respecting some loose hierarchical structuring in sections, sub-sections, paragraphs and sentences. These systems do not attempt to understand the text (unless some expert system approach is adopted), but still allow an effective retrieval on text. In fact, as opposed to image objects, they can exactly recognize words (as ASCII patterns), on which they base their retrieval capabilities, with the possible help of thesauri to support synonyms [17]. This is possible, in case of text, because a common semantic is associated to the words used in the natural language. Hence, both DBMS and IRS approaches cannot be directly applied to the image retrieval.

In addressing the problem of image retrieval on large volumes of stored images, if we want to think of a system doing for images what DBMS and IRS do for formatted data and text, we must accept some indeterminateness, characteristic of images, and then deal with the inaccuracy introduced by this fact. We have recognized the exigency of adopting, at some level of the image recognition process, a non-boolean logic allowing some form of approximate reasoning.

Thus, we have decided to investigate an approach based on the mathematical theory of evidence, for the image recognition and description part, and on the probabilistic theory, for the retrieval of the images previously analyzed. This combined approach allows the representation of different degrees of similarity between objects, the recognition of images, and contained objects, with certain belief, and the definition of classes of images/objects with unsharp boundaries.

We must stress the point that the real goal of this image analysis process is not to attempt any deep image understanding, but is to support the image retrieval process. In fact, the system allows the user to query the images, already analyzed and stored, giving some specification of their content. The image query processing is based on special access structures (i.e. image indices) which are generated when the image analysis is performed.

It has been shown that the major difficulty, related to image retrieval by content, is to define and interpret the content of images given the large variety of objects that can be contained and the complexity of the relationships among them.

For these reasons, it is firstly necessary to choose a class of images to handle (i.e. application domain) and then to build an efficient image classification system that allows, after a representation and analysis phase, the description of images in terms of the objects they contain. Moreover, in order to facilitate and to speed up retrieval operations, an efficient image database is needed, in which image descriptions are organized and suitable access structures are built.

In the image sub-system of MULTOS we can identify four different tasks:

1) The purpose of this task is to describe the characteristics of the application domains of the images to be classified and retrieved. This description is in terms of patterns and rules that will be exploited in the image analysis phase (Task 2) to correctly "understand" the images entered into the system.

2) The purpose of this task is to analyze the images entered into the system using the definitional information (particular of the application domain) specified in Task 1. This process can be extremely complex and time consuming, mainly for images entered as bit-maps. For this reason we want to split this task in two sequential phases:

   2A) The low-level image analysis accepts an image constituted by a bit-map and recognizes the basic objects in it, their relative positions and the associated distinguishing attributes. Since the number of these basic elements, obtained from the image scanning process, can be very large in a single image, a very efficient approach is required. It is not possible to adopt a rule system, based on a generalized inference mechanism, since the computational complexity would be exponential. We need instead a polynomial computation complexity, even if we have to pay this with a description system less rich in semantic content. For this reason we have adopted an approach based on the Attributed Relational Graphs.

   2B) The high-level image analysis starts from the basic objects obtained in the previous phase and then applies a body of rules for the recursive recognition of the objects contained in the image. In this phase a generalized inference mechanism is used. The computational complexity is acceptable now, since fewer higher level objects are present in the image. The result is the interpretation of the image, in terms of contained objects with associated belief and plausibility, and their attributes. The theory of evidence is used for the extraction of the belief intervals of the different interpretations of the objects from all the recognitions of the objects in the image.

3) The information obtained in Task 2 (i.e. interpretation of the images and information about the contained objects) is then used to generate access structures on image content, which can be used for efficient image retrieval (in Task 4). Access structures are mainly indices on the objects contained in all the image, with the associated attributes, and clusters about the interpretations of the images.

4) The purpose of this task is to accept queries on image content and to retrieve the images, stored in the system, which satisfy in some degree the query specification. For the image query processing task, the access structures generated in Task 3 are used.

6. Uncertainty Management.

We have presented a typical query that can be asked when complex documents are stored. Then, the query has been translated ini the MULTOS query language, but it is obvious that it does not allow to express exactly that query.

The extensions presented in sections 4 and 5 allow to obtain, as a result of the retrieval process a list of documents ranked in order of the relevance to the query.

We are now going to describe how the Query Language can be adopted in order to make the user able to express his uncertainty on part of the query components.

To each query component is associated an *importance* value, while a *preference* value is associated to each attribute value.

The importance can be expressed by the user during query formulation and it can assume the values *im* = {HIGH, MEDIUM, LOW} while the term preference has different evaluations for exact match and partial match terms. Exact match query terms contain restrictions associated to conceptual-components (e.g. product-price = 5000$) while text will be the primary partial match attribute also if figures, images [4] and voice can be other candidates for partial matching.

For exact match attributes the user expresses his preference on different term values; two possible preference levels are possible {PREFERRED, ACCEPTABLE}.
For partial match terms such as text and image components the evaluation of pm_i is based on the methods presented in the previous sections.

The previous query can now be extended with conditions on the content (in terms of contained objects) of the image required in the documents to be retrieved:

```
FIND 100 DOCUMENTS SCOPE = PC-documentation TYPE = offer-letter WHERE
(product-price BETWEEN (4000,4500) ACCEPTABLE,
               BETWEEN (4501,5500) PREFERRED,
               BETWEEN (5501,6000) ACCEPTABLE) HIGH
(company-name = Olivetti) HIGH
(TEXT CONTAINS good ergonomics) LOW
(IMAGE MATCHES
               screen HIGH
               keyboard HIGH
               AT LEAST 2 floppy_driver LOW) HIGH
```

The query allows to express that the *product-price* is approximately equal to $ 5,000 and that the values around $ 5,000 are preferred; furthermore it allows to express that the condition on the *company-name* is more important than the other.
The *preference* values associated to each attribute value and the *importance* values associated to each attribute are used for document ranking.

After a retrieval, the set of documents are presented to the user as an ordered list. The ordering is given by a score associated to each document. The score is obtained as a combination of preference values, partial matching uncertainty values and importance values associated to each predicate in the specified query based on probabilistic models.
If the preference for a given document of query attribute i is pr_i and the importance of the attribute is im_i than the document score is $\sum_{i=1}^{k} pr_i \times im_i$ where k is the number of query terms.

The query execution strategy presented in chapter 3 has to be modified to provide ranking based on importance and preference values. A possible simple strategy requires to group together all the atomic queries where exact matching on formatted data is required (l_1) and the queries on partial match (l_2). Than l_1 and l_2 are executed separately and the scores for retrieved documents calculated.

There is however a lot of room for query optimization both trying to reduce the number of accesses to formatted data depending on the preference and importance values and trying to limit the access to the signature file only for a limited number of documents. This work is currently ongoing and some initial results are available.

7. Conclusions.

The problem of managing uncertainty for attaining better results in retrieving multimedia documents is discussed in this paper. We presented the techniques used in managing the uncertainty in evaluating the significance of image and text component in relation to the user's query. The combined use of these techniques allows the ranking of the retrieved documents in order of relevance to the query. This means a better understanding by the system of the user needs.

The techniques presented for managing uncertainty in the document retrieval process have been studied in the context of project MULTOS. We have also presented the essential characteristics of the MULTOS prototype system. The techniques applied in this system are intended to support powerful content based retrieval and efficient storage of large numbers of multimedia documents.

A first Multos prototype has been developed since September 1987: it handles only textual documents. The implementation of the final prototype is in progress (the delivery is expected for June 1989). An intermediate prototype that allows to create, classify, store and retrieve multimedia documents is available.

Furthermore, a MULTOS prototype system that supports management of user uncertainty, incorporating the ideas presented in this paper, is expected for the end of 1989.

References

1. N.J. Belkin and W.B. Croft, "Retrieval Techniques," *Annual Review of Information Science and Technology (ARIST)*, vol. 22, pp. 109-145, 1987.
2. E. Bertino, A. Converti, H. Eirund, K. Kreplin, F. Rabitti, P. Savino , and C. Thanos , "MULTOS - A filing server for Multimedia Documents," *Proc. of the 1st EURINFO '88 Conf.*, pp. 435-442 , 1988.
3. S. Christodoulakis and C. Faloutsos, "Design Consideration for a Message File Server ," *IEEE Transactions on Software Engineering* , vol. SE-10 , no. 2 , pp. 201-210 , 1984 .
4. P. Conti and F. Rabitti , "Retrieval of Multi-Media Document Images in Multos," *Fourth Esprit Conference*, pp. 1389-1412, North Holland, New York-Amsterdam, 1987.
5. W.B. Croft and P. Savino, "Implementing Ranking Strategies using Text Signatures," *ACM Transactions on Office Information Systems*, vol. 6, no. 1, pp. 41-62, 1988.
6. W. B. Croft and R. Krovetz, "Interactive retrieval of office documents," *Proc. Conf. on Office Information Systems*, pp. 228-235, 1988.
7. H. Eirund and K. Kreplin, "Knowledge based document classification supporting integrated document handling," *Proc. Conf. on Office Information Systems*, pp. 189-196, 1988.

8. W. Horak and G. Kronert, "An Interactively Formatting Document Editor Based on the Standardised Office Document Architecture," *Proc. of IFIP Conf. OFFICE SYSTEMS: Methods and Tools*, pp. 287-300, 1986.

9. M.D. McIlroy, "Developement of a Spelling List," *IEEE Trans. Commun.*, vol. COM-30, no. 1, pp. 91-99, 1982.

10. J.M. Morrissey and C.J. Van Rijsbergen, "A Formal Treatment of Missing and Imprecise Information," *Proc. of the 10th Annual International ACMSIGIR Conference on Research and Developement in Information Retrieval*, pp. 149-156, 1987.

11. ODA, *Office Document Architecture*, ECMA-101, 1985.

12. F. Rabitti and J. Zizka, "Evaluation of Access Methods to Text Documents in Office Systems," *Proc. 3rd Joint ACM-BCS Symposium on Research and Developement in Information Retrieval*, 1984.

13. C.J. Van Rijsbergen, *Information Retrieval*, McGraw-Hill, London: Butterworths, 1979. 2nd Edition

14. C.J. Van-Rijsbergen, "A Non-Classical Logic for Information Retrieval," *Computer Journal*, no. 29, pp. 481-485, 1986.

15. S.E. Robertson, "The Probability Ranking Principle in IR," *Journal of Documentation*, vol. 33, pp. 294-304, 1977.

16. G. Salton, *Automatic Information Organization and Retrieval*, McGraw-Hill, New York, 1968.

17. G. Salton and M.J. McGill, *Introduction to Modern Information Retrieval*, McGraw-Hill, New York, 1983.

(79.) Conti P., Rabitti F., Savino P., Stanchev P., "Managing Uncertainty in a Multimedia Document System", Proc. of ***Conference on Intelligent Management Systems,*** Varna, Bulgaria, 26 Sep.- 2 Oct. 1989, 177-186.