

---

## Issues of data modelling in information retrieval

M. AGOSTI

*Dipartimento di Elettronica e Informatica  
Università di Padova  
Via Gradenigo 6/a  
35131 Padova, Italy*

R. COLOTTI

*SoftLine  
Via Piovese 108  
35127 Padova, Italy*

G. GRADENIGO

*Dipartimento di Elettronica e Informatica  
Università di Padova  
Via Gradenigo 6/a  
35131 Padova, Italy*

---

### SUMMARY

This paper addresses the problem of data modelling in information retrieval. The study introduces various aspects and issues that are necessarily taken into account when designing and developing an information retrieval system. Particular attention is paid to the representation of the different types of data managed by an information retrieval application: structured and unstructured data.

A recently introduced information retrieval, data modelling approach supports the notion of a schema permitting representation of the information retrieval data on two different levels: intensional and extensional. The characteristics of this data modelling approach are presented here together with examples of its use in a working prototype.

**KEY WORDS** Data modelling in information retrieval Data representation by content Text representation by content Information retrieval model Information retrieval conceptual architecture Auxiliary data Hypertext information retrieval Hypertext capabilities

## 1 INTRODUCTION

This paper addresses the topic of data modelling in information retrieval (IR). The first part is an introduction to the basic issues related to this topic. This part is based on Reference [1] in which a non-traditional approach to data modelling in information retrieval is introduced.

The second part of the paper presents an information retrieval technique based on a two-level architecture. This method is developed to face the issues of data modelling presented in the first part. The paradigm which distinguishes this approach has been introduced in Reference [2]. An account of the characteristics of the architecture is given in References [3] and [2]. A report of the current stage of development of a model based on this technique is given in References [4] and [5].

## 2 DATA MODELLING IN INFORMATION RETRIEVAL

### 2.1 Documents and document collections

Data concerning a specific application domain can be modelled to two different abstraction levels, one *intensional* and the other *extensional*. These two levels of representation generally reflect the user's conceptualization of the 'reality' to be represented in a data management system.

---

At the *intensional level* the formal description of properties and relationships of a set of informative objects, specified in terms of the semantic or logical structure of a particular database model, is known as the schema. Design and management operations on data can be effectively carried out by using a schema, disregarding the particular implementation.

At the *extensional level*, the objects of the reality of interest described in the schema populate the database.

Data in the information retrieval environment are representations of documents. By the term *documents* we generally mean those objects which contain information of any media type: i.e. books, reports, photographs, slides, videos, audio tapes, and so on. It is important to note that at present the information related to and/or contained within these different multimedia objects is usually represented as *textual data* only. This means that present state-of-the-art IR systems manage only textual representations of documents.

Thus a collection of multimedia documents is represented by a collection of textual fragments incorporating different elements, such as author data, title, date of publication, etc., as a deterministic representation of the document; in the following, this representation is referred to as *structured data*. Other textual data represents the semantic content of each document of the collection; this data is called *unstructured data* in the following and represents the output of a non-deterministic representation process. The IR system manages these two types of representation including the connection with the original document that can be on any kind of media. The system then provides data regarding the physical position of the document or it displays (or plays) it to the user.

In this paper we address the general problems faced in the management of IR data, overlooking the physical structure and the media typology of the documents.

The IR textual data representing each document is expressed by structured data and unstructured data:

- *structured data* are all those data items which describe the deterministic nature of the document (e.g. date, author, place and year of publication);
- *unstructured data* represent the informative or semantic content of each document; unstructured data cannot be defined in a deterministic manner (e.g. keywords from a thesaurus, abstract, etc.); unstructured data are called 'unstructured' because it is not possible to manage it at a logical and physical level within a record structure with fields of fixed length, i.e. it is not possible to arrange this data according to attributes of a relation of a relational database management system.

The existence of these two kinds of data within an integrated database establishes the peculiar character of the management in information retrieval data, especially due to the presence of unstructured data.

## 2.2 Representation and management of structured data

Management of structured data is based upon a consolidated technology: common data processing applications such as accounting or planning are able to manage only structured data. The integrated collection of data generated by applications of such nature consists basically of structured data which can be designed by making use of database design methodology. Data management is generally implemented by means of a traditional database management system.

---

Both the system and the database design methodology can be based upon an operational data model, establishing a common reference framework for data representation and data management. Examples of well-known operational data models are those concerning the relational and network kinds. These models have been developed and proposed mainly for management purposes but they are powerful enough, also enhanced by means of external conceptual tools (see the entity-relationship conceptual model for the relational model), to permit representation of data according to different levels of abstraction.

### 2.3 Representation and management of unstructured data

The kind of data which is peculiar to ordinary information retrieval operations is that which represents the informative or semantic content of the document collection: i.e. unstructured data. For this reason, we now concentrate the discussion on this type of data.

It is essential to note that so-called unstructured data, in the information retrieval area, is generally not really 'unstructured', due to the fact that it often consists of a collection of terms organized into a semantic structure which is used by the information retrieval system to find the correspondence between terms in the user's query and the documents of the managed collection. Through this structure the information retrieval system extracts some of the IR documents from the IR document database in order to satisfy the user's information requirements which have been expressed in his query. The term 'unstructured' results mainly from the fact that it is inadequate to represent the semantic content of a collection of documents by means of a completely flat set of terms.

Different structures have been proposed to manage the collection of terms used by the information retrieval system in order to represent the domain of pertinence of the collection of IR documents together with the terms within the structure. The simplest structure used to represent and manage the collection of terms basically consists of an alphabetic list of significant words or terms; other more complex structures can be that of a classification scheme, a semantic network or a thesaurus.

Many operational information retrieval systems are able to manage different structures encompassing various different collections of terms in a concurrent manner. The concurrent availability of different structures permits representation of the collection of documents according to different levels of specificity and/or for different categories of users. When this is the case, the user is given the possibility to choose at run time what sort of unstructured data to use in his queries, or, when possible, to choose a combination of different sorts of unstructured data.

It is possible to introduce a classification of unstructured data due to the nature and application of this sort of data; it is therefore possible to consider two different families of unstructured data: surrogates and auxiliary data.

*Surrogates* are basically textual data which often substitute the real document and represent its content within the system; surrogates can be: abstracts, indexes of documents, prefaces, and so on. Surrogates do not make reference to any type of structure; they are chunks (fragments) of natural language processed at input time in the database by a parsing algorithm which extracts significant words. These words are collected together to form a list of words used by the information retrieval system as a collection of entry points to the documents. When the user formulates an enquiry to the system using one of these words, the system presents the user with details concerning all the documents which contain the word in the surrogate.

---

*Auxiliary data* is another different technique for representing the informative content of a document. Auxiliary data generally consists of a semantic structure incorporating a collection of words or terms which depend upon the domain of the collection of documents managed by the information retrieval application. The auxiliary data structure identifies the semantic relationships between words or terms: it is this structure which permits implementation of data retrieval operations by content in contrast to the usual data retrieval by value, currently supported by traditional data processing applications. Many different paradigms and techniques can be used in the design and management of the structure of auxiliary data: i.e. object-oriented, knowledge-based, etc. All approaches have the same basic purpose: to implement a structure which can be used to find semantically related terms; by using related terms which directly concern the user's specific information needs it is possible to launch a query to the database of the collection in order to retrieve all relevant documents.

Auxiliary data items represent the vocabulary to be consulted by the information retrieval system; this means that such a structured collection of data can exist even if the documents have not yet been inserted into the system. These auxiliary data items are associated to each document by means of an indexing process in order to represent its semantic content in the database. Each auxiliary data item is assumed to describe the document content only to a certain extent, neither completely nor uniquely. Various different sets of auxiliary data items may be assigned to the same document; in fact, the description of the content of a specific document does not arise from a deterministic process as in the case of surrogate data. Furthermore, the content can either be described in different ways by several different people or the different descriptions simply reflect different users' requirements for information.

Some examples of structured data can be the values of possible attributes of a document, e.g. AUTHOR, TITLE, and DATE. Of the same document, surrogates can: ABSTRACT, CONTENT-INDEX, PREFACE. And EXAMPLES OF auxiliary data are a classification system and a thesaurus.

An example of the representation of an IR document is reported in Figure 1. This representation is ready to be stored and managed by an operational information retrieval system. This IR document has been extracted from the INSPEC on-line database by the ESA-QUEST information retrieval system. The various different parts of the document have been named according to the terminology previously introduced. This figure gives a clearer understanding of what we mean by structured and unstructured data and the differences existing between these two kinds of data.

The information retrieval models which have been proposed for application in the information retrieval area (see References [6] and [7] for a general introduction to the subject) differ considerably from one another, basically due to the different nature and structure of the auxiliary data used in the operations of the IR system based on these models.

Information retrieval models which permit an exact match between the user's query and the unstructured data are capable of managing surrogates only. These models make use of a representation of the document informative content actually given by the surrogates or, more simply, of a list of words associated to the documents. It is important to note that a list of alphabetically ordered terms contains no semantic relationship between its component words, thereby it merely represents a flat and semantically poor structure. The

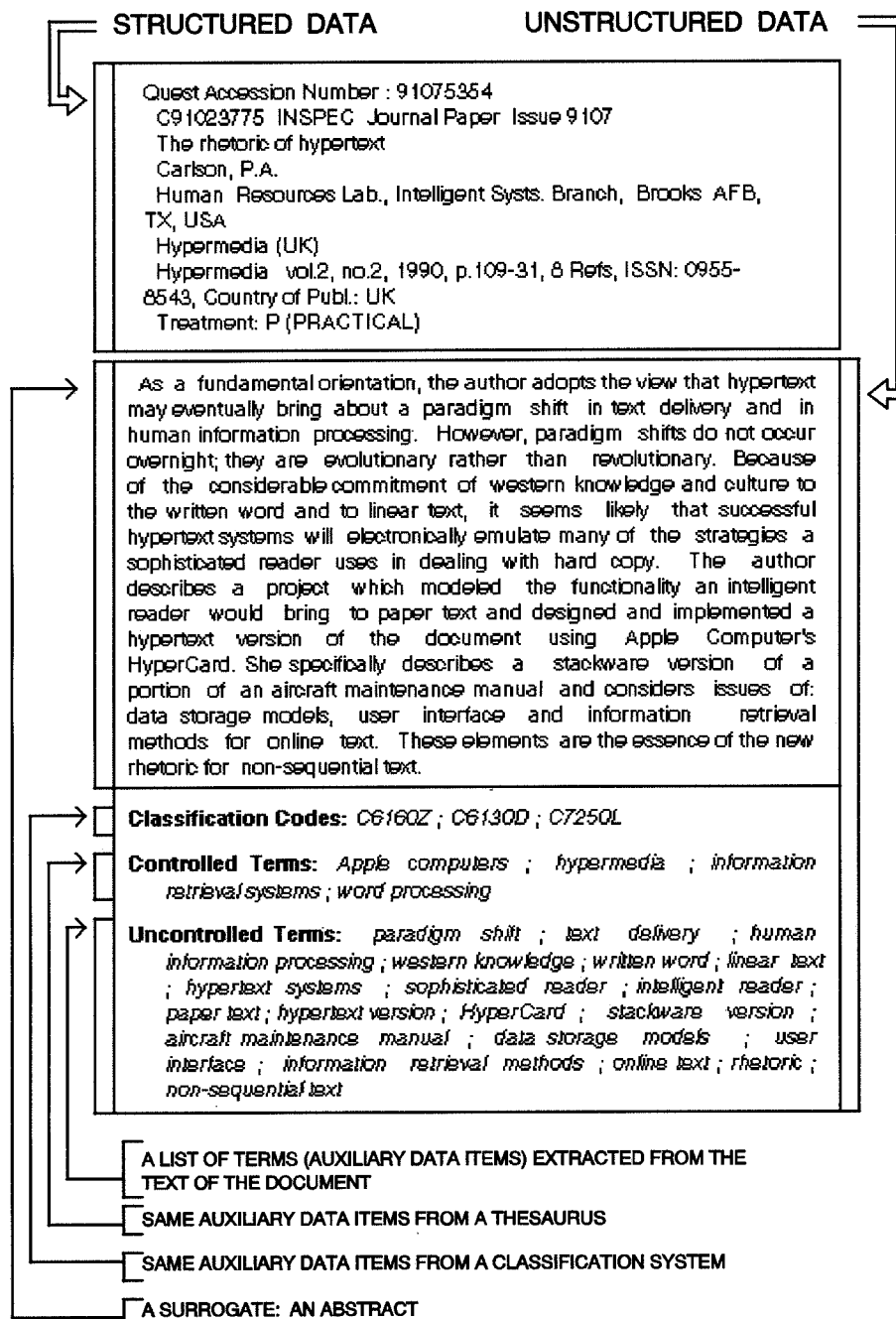


Figure 1. A document extracted from the INSPEC database by the ESAIQUEST information retrieval system

---

most well-known model among the exact-match based information retrieval models is that based upon Boolean algebra.

Information retrieval models based on a partial match retrieval technique are also capable of operating an exact match. These normally use a specific auxiliary data structure. Depending on the auxiliary data structure identifying the model, different types of information retrieval systems can be implemented. One of these models is the probabilistic one [7]. Some other models based on hypertext techniques are currently being developed; two such proposals have been presented in References [8] and [9].

### 3 A NEW APPROACH TO DATA MODELLING

#### 3.1 Usefulness of an explicit auxiliary data structure

Information retrieval systems, disregarding the implemented model, do not normally foresee a schema which intensionally represents the information stored in the database. This lack of a formal description of the database content deprives the user of an explicit reference upon which to formulate a query. It is not, however, seen as a serious drawback when the query merely involves information of deterministic nature (e.g. 'find all reports with date of creation = 1989'), but it becomes extremely serious when the information sought concerns the document's informative content (e.g. 'find all reports regarding retrieval techniques'). On most information retrieval systems currently available, a single indexing term or a list of these terms used in a document's informative content representation are normally displayed to the user, but these terms cannot be englobed into a semantic context in which the meaning of each term is defined by its relationship to other terms.

A schema of the concepts and of the relationships existing between concepts which describe the informative contents of the document collection and produced for a specific information retrieval application domain, if made explicit, would provide the user with a *frame of reference in the query formulation process*. Users of information retrieval systems normally look for documents merely to retrieve the information stored in the document collection.

A browsing feature which could permit the user to wander through the structure of concepts representing the informative content of the document collection would enhance the user-system communication and consequently the effectiveness of the system. If the user is given the possibility to navigate through the semantic structure of concepts (the auxiliary data structure), a powerful interaction with the information stored in the document collection managed by the information retrieval system is made possible.

In the authors' opinion an important requirement for an information retrieval schema would be the availability of a model which could allow a clear expression of the conceptual structure of the auxiliary data items as well as of the complex relationships existing between them. Due to the important role played by the auxiliary data in determining the effectiveness of an information retrieval system, the need for a conceptual tool supporting the designer of the information retrieval database and of the conceptual structure of the auxiliary data becomes essential. Such a model would need to be automatically processable in order to allow the representation of the semantic structure of the auxiliary data to be dynamically managed and updated. An automated management scheme of this model would also permit the information retrieval user to make effectively interactive use of the auxiliary data.

### 3.2 The conceptual modelling paradigm

Specification of a semantic structure of auxiliary data involves devising a conceptual modelling paradigm by which to support the abstraction mechanisms necessary for the semantic structuring of the indexing terms. Three abstraction mechanisms are generally considered as being essential: the classification mechanism, the generalization/specialization mechanism and the aggregation mechanism. These three

mechanisms are hereby illustrated. Major references for the formulation of a conceptual modelling paradigm are [10,11] in the database area and [12] in the hypertext area.

The classification mechanism is one of fundamental and very intuitive nature which permits definition of a class selected from a set of objects with common properties. Each single object is a defined instance of the class; a class and its instance are related by means of an 'instance-of' relationship. This relationship is implemented by associating the term identifying the class to the single document. In an information retrieval system, the class definition supplies the end user with a description of the database informative content. The classes delimit the specific area in which the system is recognizable and define the vocabulary to be used to access the information. For classification of information retrieval data it is essential to use a polythetic classification mechanism [7]; in this way each component of a class receives only a portion of all the attributes possessed by the members of the same class; hence no attribute is both necessary and sufficient to determine the property of any one element belonging to a specific class. For example, a scientific paper usually concerns some specific topic and therefore needs to be indexed by several terms, each term identifying a class to which it should belong.

The generalization/specialization mechanism [10] simply relates a set to its subsets or a class to its subclasses. The representation of this mechanism is usually expressed by a 'subset-of' or 'subclass-of' relationship. An example can be given by examining the relationship existing between documents belonging to a certain collection on the main subject of computer science and the subclasses of documents that can be created by distinguishing the documents within the collection into classes of different computer science subtopics. To organize and provide associative access to a real document collection it is rather commonplace to use some kind of classification scheme. By means of a classification scheme each document of the collection is classified into a subclass which groups together all documents concerning a specific subject. Thus a classification scheme can be seen as a mechanism by which a hierarchical tree structure of subclasses of documents can be built. The collection of such documents is therefore transformed into a well-organized, tree-structured order.

The aggregation mechanism [11] establishes a relationship between objects of the same, or of different nature by transforming a relationship connecting several different objects into a single object of higher level. This new object can have definite, unique characteristics of its own. An aggregation abstraction mechanism is generally used to model the user's perceived relationship between concepts. This mechanism can, for example, establish a relationship between the terms 'information retrieval system' and 'retrieval technique'.

### 3.3 A new approach

As demonstrated, the complexity of the task of data modelling in information retrieval systems is mainly due to the complex nature of the relationship existing between separate

---

auxiliary data items. In fact, auxiliary data generally gives a description of the document's informative contents, but the meaning of an auxiliary data item can become fully defined only in the context of the semantic relationship existing between this item and other terms.

The difficulties encountered in the *design* of an *information retrieval system* capable of establishing an *explicit distinction between documents and auxiliary data* are hereby presented together with a proposed solution. An architecture and a applicable approach based upon it are briefly presented in order to be used in the following as an example in the presentation of the various methodological aspects. For this reason the information retrieval approach is introduced merely to demonstrate in what way the user can interact with the system and the new capabilities he can use, being out of the scope of this paper to provide a formal presentation of the approach.

The information retrieval approach makes the conceptual structure of the indexing terms used in the application explicitly available to the user. It is based on a two-level architecture, which accommodates the two main parts of a database managed by the information retrieval system, i.e. the collection of documents and the auxiliary data: structure + collection of words. In order to make the presence of these two sections of the information resource managed by the system explicit for the user's understanding, an architecture which permits work on two different levels of abstraction has been taken into consideration:

1. the level of documents (e.g. full text documents);
2. the level of semantically related concepts: this is the plane of abstraction where indexing terms are arranged together with their structure. The objects of this level are a direct result of the application of the classification abstraction mechanism to the objects of the first level.

The connection between the two levels establishes a correspondence between the concepts and the documents which are instances of these concepts (see Figure 2). This technique helps to make up a network set up by two communicating sub-networks which can be represented and managed by a system having hypertext capabilities.

The abstraction mechanisms referred to above turn out to be satisfactory for the modelling and organization aspects of the representations of information which may be contained within the documents of the collection.

The two-level architecture permits an explicit representation of the conceptual modelling of the information retrieval data, making it possible to produce a schema of the concepts describing the informative content of the document collection being managed by the system.

The independent nature defined by the particular architecture between its two levels; i.e. hyperconcept and hyperdocument, allows us to use this solution even for systems already operational.

The experimental experiences that have been set up to validate the architecture and the approach were directed mainly towards two specific aims:

- usage of the architecture as an interface of an already consolidated information retrieval environment;
- design and development of an information retrieval system based on the proposed architecture.



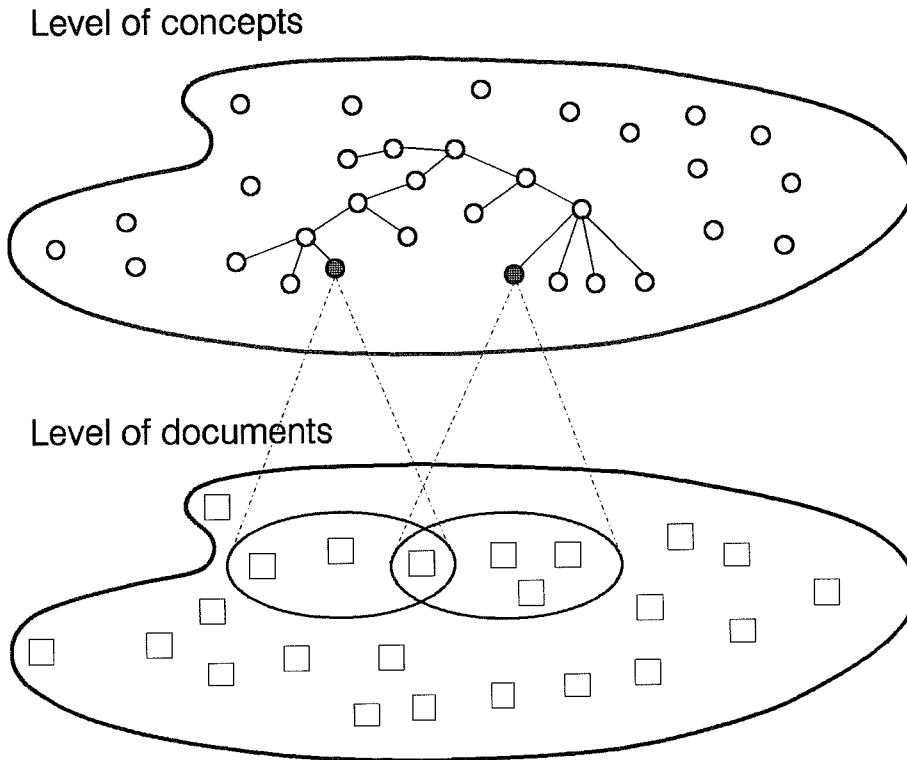


Figure 2. The two-level conceptual architecture

The first experimental setting has been an *on-line information retrieval service* which manages large bibliographic collections: the Information Retrieval Service of the European Space Agency (ESA/IRS). In this type of setting, the architecture and the basic features of the model have been used to design and implement a conceptual interface to very large on-line bibliographic collections [4,13].

The interface thus developed, called HYPERLINE, works as a conceptual reference tool for the final user of an available information retrieval system, playing a communicative role between the user and the system [4,13]. The first level of the architecture, the document collection, is implemented and managed by the ESA-QUEST information retrieval system; the second level of the architecture has been designed and implemented as an interface which permits interactive use of the auxiliary data concept structure and the relationships which exist between the level of documents and that of the concepts and vice versa.

The second experimental setting is a *personal computing environment*. This second direction foresees the development of an information retrieval system totally based on the new approach [5].

Personal computer users usually pay a special attention to problems related to user interfacing, sometimes giving up somewhat on performance requirements. Those who make use of document collections on personal computers normally interact directly with the document collection itself and formulate queries in an interactive manner. The tool which provides access to the information contained within this environment must therefore

---

be capable of balancing the handling efficiency with the most natural usage of the system. This approach seems to offer the necessary opportunities, due to the fact that the developed system manages the document collection, the auxiliary data structures and the relationships between them in a direct fashion, thanks to the hypertextual capabilities it offers.

The prototype thus developed, called HyperLaw, manages a collection of full-text legal documents. Motivations and an initial version of the prototype have been presented in Reference [14]. Characteristics of a consolidated version of the prototype are presented in References [15] and [16].

In the following sections the different aspects concerning user interaction and available capabilities are presented, analysing different aspects at each level of the architecture and in the relationship between the two levels. The examples and figures introduced make use of the documents and auxiliary data of the HyperLaw prototype.

### 3.3.1 *The level of documents*

The document collection is represented and managed at the first level of the architecture (see Figure 3). Each object of the collection is a physical entity directly modelled to a 1 to 1 correspondence with an object of the representation. Each object has its own identity and status; the identity of the object is independent of the manner in which it is represented or structured and of the values it may assume. The identity makes each object distinct from all the other objects of the collection.






At present this approach supports only one representation of an IR object; in the future it will need to be extended in order to support different representations of the same object, thus permitting an effective management of *multimedia documents*.

The use of this approach for the representation of the most common information retrieval objects, i.e. textual documents, allows the employment of:

- a set of structured data which represents the different deterministic properties of the object (e.g. date of publication, title, list of authors);
- a text fragment or summary of the informative content of a document;
- connections to other documents which are related to a specific one; the connections establish relationships between documents according to their informative content (e.g. documents concerning the same subject), or belonging to the same series (e.g. the series of proceedings of an annual conference);
- connections to the auxiliary data items (i.e. indexing terms) which represent the informative content of the object; the auxiliary data items and their structure are represented and managed at the second level of the architecture.

Representation is often managed with a part of, or the whole of the document. At the extensional level of the architecture, the collection of document objects forms the '*hyperdocument*', that is a lattice structure in which the document base is represented and organized. This means that each node of the hyperdocument is an informative item consisting of the representation of a document together with the full text or simply a part of the document itself.

In this way, a complete document is represented by a node or by a set of nodes; when the document is represented by a set of nodes these nodes are connected by means of *structural links*. A structural link therefore serves to structure the parts of the same document in a hierarchical structure.

|  |  |
|--|--|
| <p>AUTORE cicala mario<br/>         TITOLO la legislazione contro il rumore industriale ed ambientale in italia relazione per il congresso internazionale medico: l' uomo e il rumore: problemi biologici, audiologici, industriali e giuridici, torino, 7-10 giugno 1975<br/>         PERIODICO Giur. merito, an. 7 (1975), fasc. 4-5, pt. 4, pag. 212-228<br/>         (Bibliografia: a pie' di pagina o nel corpo del testo)</p>  | <br><br><br>Go back<br><br><br>References<br><br><br>Mark<br><br> |
| <p>RIASSUNTO l' a. individua 4 partizioni fondamentali nell' ambito della legislazione italiana sui rumori. la prima, disciplinata dal codice civile, trova nella tollerabilita' il criterio scriminante che non puo' essere superato e che permette una larga discrezionalita'. la seconda partizione e' regolata dal codice penale che prevede una doppia ipotesi: quella di chi disturbi le occupazioni e il riposo delle persone e quella di chi esercita una professione rumorosa contro le disposizioni di legge e le prescrizioni dell' autorita'. l' a. nota che anche l' esercizio di un diritto costituzionale quale la liberta' di pensiero deve essere temperato con le esigenze della normale convivenza. terzo settore e' quello regolato dal diritto del lavoro in cui al di la' del generico obbligo di contenere l' intensita' del rumore nel</p> |  |

Level of documents

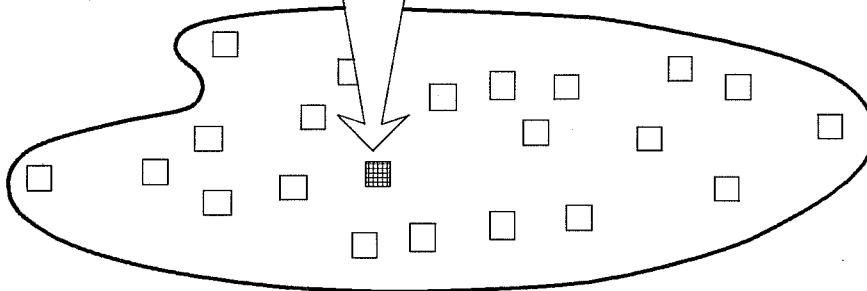


Figure 3. The first level: the collection of documents and the representation of a specific document within the collection

Two different documents related by means of a citation or reference are connected through a *reference link*. Reference links are the other kind of links which can be used for structuring the hyperdocument.

In using this approach for structuring a collection of law documents the structural links can be used, for instance, to arrange the different articles of a law, and the reference links to connect a part of a new law to the part of another law that is cited by the new one.

*Structural links* permit proper organization of the nodes because these actually represent

---

and implement the physical relationships existing between parts of documents. Structural links reflect the hierarchical structure existing in the original documents. These structural links are obtained by connecting a father node to its offspring in order to form a branching diagram within the global diagram of the hyperdocument. These links are predefined and the user cannot modify them but simply use them to browse the database.

*Reference links* permit representation of the semantic relationships existing between informative contents of nodes. Reference links can also be handled by the final user according to his information interests in linking of documents.

The resulting hyperdocument consists of a network of structural links united together with a network of reference links. It is important to note that the directional semantics of a link is considered as being significant. This means that the user may choose to follow along one path or another even in consideration of the direction of the references present within the semantic units. The significance of this facility is evident if we consider the difference between the semantics used for a text being referred to, or for one text referring to another.

The approach foresees the introduction of a few functions to be used for active user-system interaction operations allowing an effective consultation of the document base. These functions are explained briefly below. This approach supports navigability through the document collection by means of the structure built up of nodes and links. Due to the fact that specific cross-references are often present among the documents of the collection, a system based on this approach must explicitly be able to support navigability through these connections. Furthermore, reading one item of information stimulates request for other information in further depth. The implementation of a hypertext network between the various information items permits their direct consultation.

To reduce the common problems connected to disorientation and knowledge overload (this is a major problem identified in the use of hypertext systems, see, for example, Reference [17]) which face the user during use of a hyperdocument, a simple searching technique for detection of text strings located within the full-text information items has been foreseen. In fact the opportunity of locating, with a certain approximation, the whereabouts of some nodes and to use them as starting points for one's own queries has been considered an important point to ponder. It appeared to be unnecessary to include a particularly sophisticated search function into this approach due to the presence of the second level of the architecture. Using the auxiliary data structure, the user can directly and actively formulate a query by which the system will go about locating the documents he is searching for.

### 3.3.2 *The level of concepts*

The *auxiliary data structure* is represented and managed at the intensional level of the conceptual architecture. The conceptual tool necessary for designing and managing the auxiliary data and its structure must be able to support polythetical classification. Since we are at the level of concepts, each auxiliary data item is considered as a concept and represented as a class. The instances of each class are the document objects which concern the specific concept expressed by the class term. Hence a class from the extensional point of view is represented by a set of documents and from the intensional point of view may be seen as a single concept in a context of related concepts, i.e. in the auxiliary data structure, which states its meaning. In fact, each concept is fully defined by its context in the auxiliary data structure.

---

In this approach, it is possible to use the multiple inheritance mechanism which permits a natural property of information retrieval to be implemented; an information retrieval object (i.e. a full text document) can be related to many other different concepts, thereby the object can belong to several different classes.

As previously stated, each class can be considered under two different levels of abstraction: as a set of instances and as a whole entity. Thus the properties of the class can be subdivided into:

- the properties of the class as a concept, that is the actual properties of the auxiliary data item;
- the properties of the objects of the class; i.e. properties of the documents belonging to the class identified by the auxiliary data item.

The level of concepts arises from the application of the classification abstraction mechanism to the documents representing the objects of the first level. Among the various different structures that have been mentioned in Section 2.3 and that can be used to manage the collection of terms to be represented by the auxiliary data structure, HYPERLINE in particular makes use of a thesaurus [4,13] and HyperLaw uses a classification scheme, i.e. a tree of concepts, which can be represented through an IS\_A hierarchy (see Figures 4 and 5).

From a user's point of view, the structure of concepts at this level plays the role of a semantic interface to the collection of documents. We have named this structure '*hyperconcept*', that is, the parallel structure whose task is to form a sort of semantic schema of concepts used to describe the contents of the hyperdocument, or the collection of document objects.

Each node of the hyperconcept represents a concept or a particular aspect of a concept pertaining to the information content of any documents which may be present within the hyperdocument. Nodes are related by links which describe the semantic relationships existing between the concepts represented by them. A different type of link is used to represent the different abstraction mechanisms (see Section 3.2) by which concepts are conceptually organized.

The user can navigate through the auxiliary data structure of this level. This navigation represents a new technique for query formulation, because this sort of query establishes a semantic path running through the auxiliary data structure. The semantic structure used in HyperLaw for the representation of the content of documents is a classification scheme; a graphical example of navigation through the semantic structure and the different windows that have been opened to show to the user auxiliary data items encountered during navigation is presented in Figure 4. The use of simpler structures (i.e. classification schemes) or of more complex nature (i.e. thesauri) has no essential significance in the making of this mechanism.

### 3.3.3 *Inter-level relationships*

The semantic link existing between two different second level class objects corresponds to a 'set' relationship between corresponding class instances of the first level. Depending on the type of link existing between class objects of the second level, different kinds of set relationships are established between the extension sets of the first level.

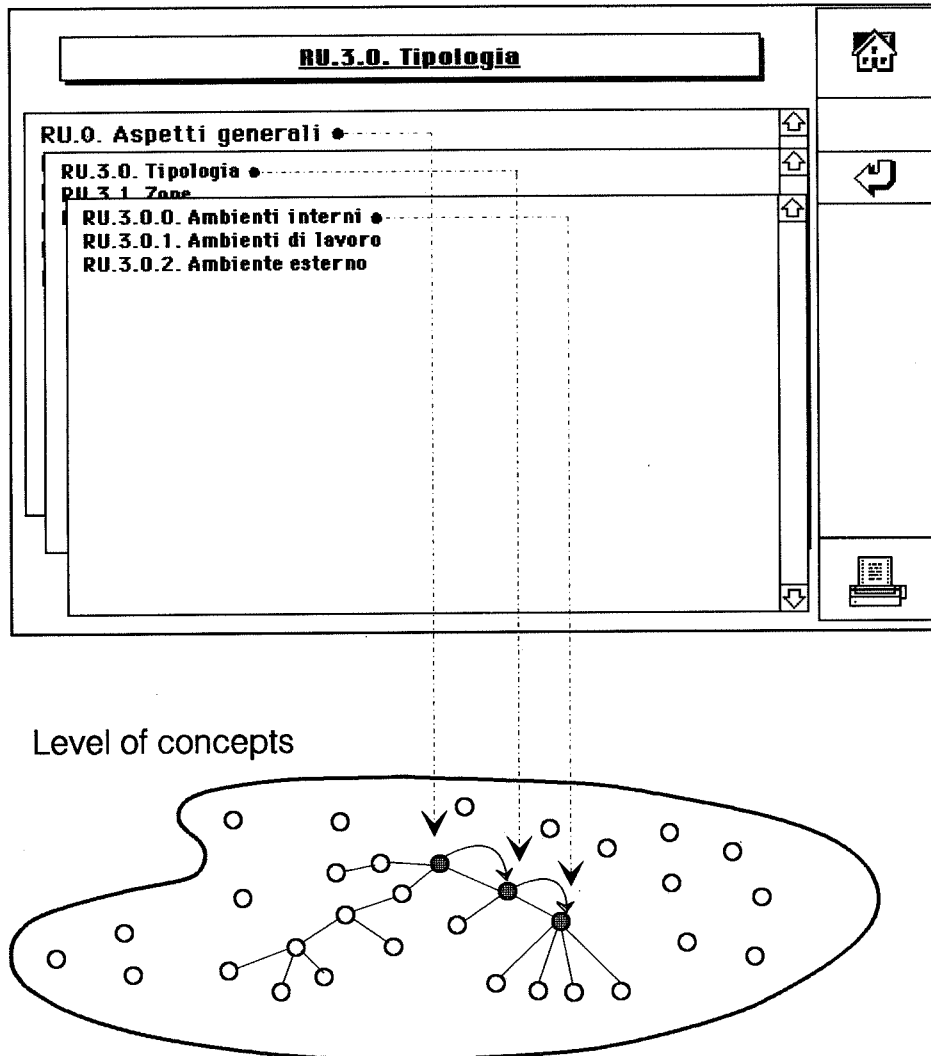
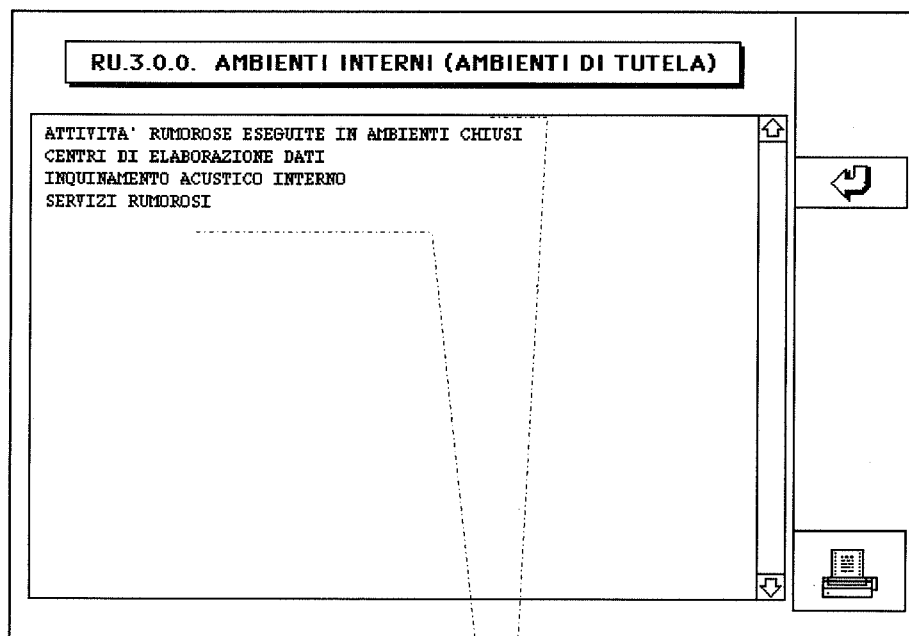


Figure 4. The second level: a graphical example of navigation through the semantic structure and different windows that have been opened to show to the user auxiliary data items encountered during navigation

This approach supports shifting operations between the two levels. In this way, it is at all times possible for the user to skip from the hyperdocument to the hyperconcept and vice versa. This feature is very important for the freedom it offers the user in actively refining his query. By shifting across the two levels, the user can in fact move towards semantic contexts that were not previously considered.

If a classification scheme of a strictly hierarchical nature is used, the semantic terms which correspond to the terminal 'leaves' of this structure represent the actual access points to the documents. If a specific term is chosen by the user at the level of concepts, the choice activates the term that gives access to an intermediate node which permits discrimination



Level of concepts

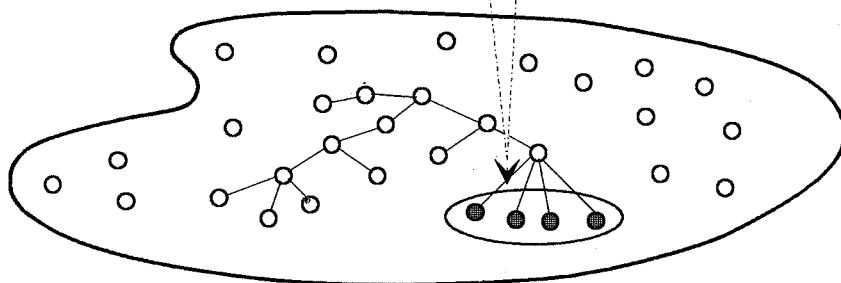
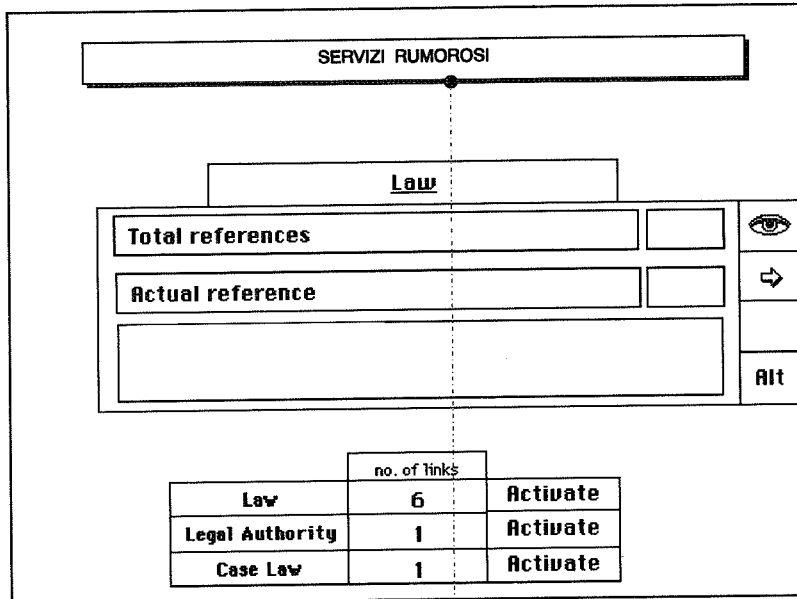


Figure 5. Graphical representation and content of leaves of a branch of the classification scheme

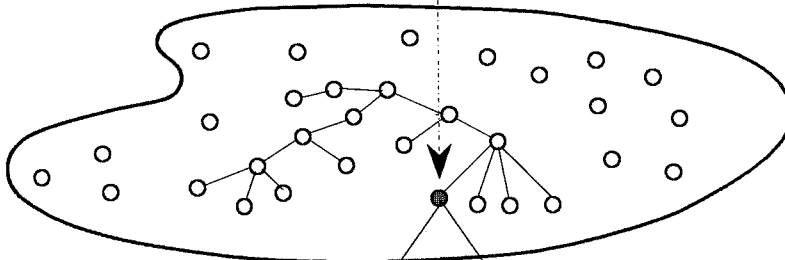
of access only to the documents semantically represented by that specific term; this passage is depicted in Figure 6 where the term 'servizi rumorosi' has been chosen and the system shows the user the number of related documents; the documents are divided in three subsets depending on the specific type of documents, namely: law, legal authority, and case law.

The user can achieve separate access to the different types of documents present in the level of documents. The cardinality of each subset of the set of documents is given (see Figure 6). If the user wishes to see the law documents, he simply chooses the 'law' set; the initial response of the system is a short reference (see Figure 7); after this short presentation, the user can see the complete document or the different documents of the set. It is possible to navigate through the documents, or through a specific subset of documents.

From any one of the hyperdocument nodes (level of documents) the user may skip



Level of concepts



Level of documents

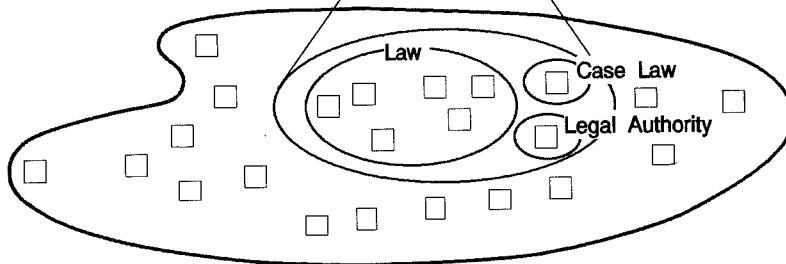
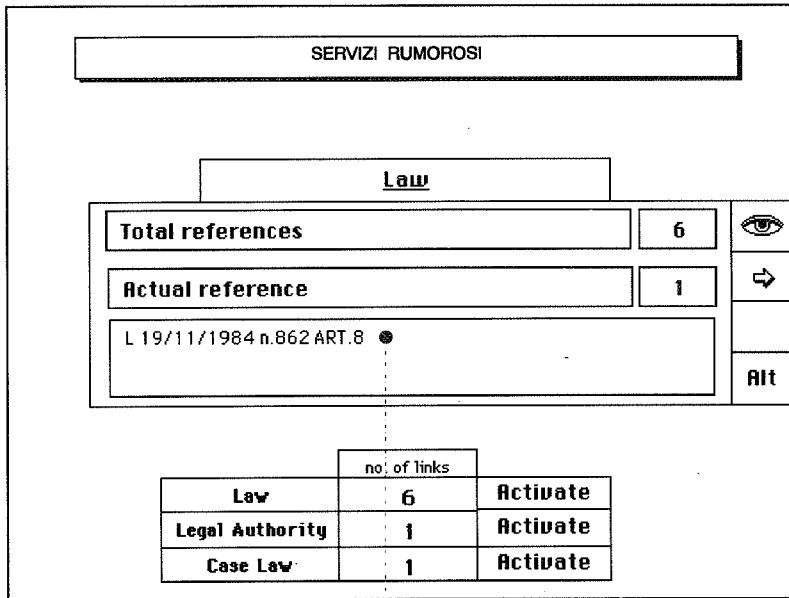
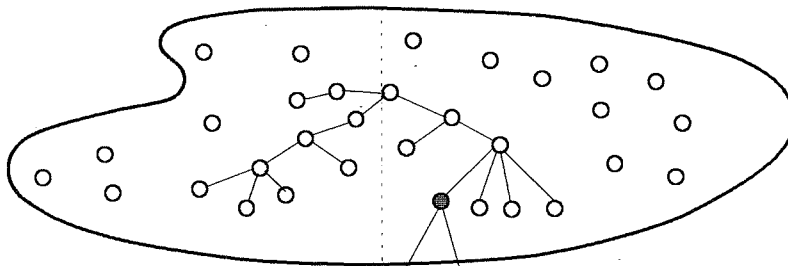


Figure 6. Example of a connection between a node of the first level and related documents of the second level; different types of documents are presented to the user





Level of concepts



Level of documents

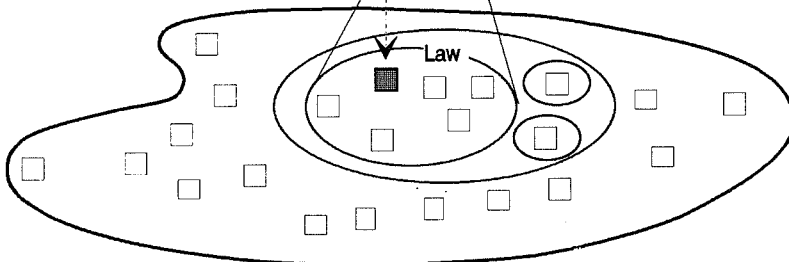


Figure 7. Example of an intermediate subsetting operation based on the previously retrieved documents; the user selected the 'law' set of documents

---

directly to the hyperconcept (level of concepts) by means of the active or passive link display functions of the node itself.

#### 4 CONCLUSIONS

The independent nature of the two levels of the system's architecture allows us to take a step further, that is, it offers us the opportunity to construct different and distinct hyperconcepts upon the same hyperdocument. In this way, it is possible to obtain different semantic descriptions of the same document collection, that is, different semantic views for different categories of users. This feature is quite significant, because a user specialized in a specific field tends to use a different terminology compared to the language used by a normal user. This means that we are given the opportunity to construct different access mechanisms allowing for different types of user interaction according to the different access requirements of the various categories of users.

This paper has disclosed the fact that the central issue concerning the field of information retrieval is that regarding the modelling of information retrieval data. An approach to the design and representation of this data has been introduced showing the purpose and the reason for the introduction of a two-level architecture.

#### ACKNOWLEDGEMENTS

The authors wish to thank Pier Giorgio Marchetti of ESA/IRS, Frascati, Italy, for the useful discussions on aspects of the work on which this paper is based.

The work of Maristella Agosti and Girolamo Gradenigo has been partly supported by the Italian National Research Council (CNR) under the project 'Sistemi informatici e calcolo parallelo—P5: Linea di Ricerca Coordinata MULTIDATA'.

#### REFERENCES

1. M. Agosti, 'Basi di dati testuali', in *Proc. of 'Sistemi di basi di dati: la prossima generazione'*, pp. 149–174, Milano, Italy (1991). AICA.
2. M. Agosti, F. Crestani, G. Gradenigo, and P. Mattiello, 'An approach for the conceptual modelling of IR auxiliary data', in *Proc. 9th Annual IEEE International Phoenix Conference on Computers and Communications*, pp. 500–505, Scottsdale, Arizona (March 1990).
3. M. Agosti, G. Gradenigo, and P. Mattiello, 'The hypertext as an effective information retrieval tool for the final user', in *Pre-proceedings of the 3rd Int. Conf. on Logics, Informatics and Law, Vol. I*, ed., A. A. Martino, pp. 1–19, Firenze (1989).
4. M. Agosti, G. Gradenigo, and P. G. Marchetti, 'Architecture and functions for a conceptual interface to very large online bibliographic collections', in *Proc. of RIAO '91 Conf. 'Intelligent Text and Image Handling'*, pp. 2–24 (Vol. 1), Barcelona, Spain (April 1991).
5. M. Agosti, R. Colotti, and G. Gradenigo, 'A two-level hypertext retrieval model for legal data', in *Proc. 14th ACM-SIGIR Int. Conf. on Research and Development in Information Retrieval*, eds., A. Bookstein, Y. Chiramella, G. Salton, and V. V. Raghavan, pp. 316–325, Chicago, USA (1991).
6. G. Salton and M. J. McGill, *Introduction to Modern Information Retrieval*, McGraw-Hill (1983).
7. C. J. van Rijsbergen, *Information Retrieval* (2nd edn), Butterworths, London (1979).
8. W. B. Croft and H. Turtle, 'A retrieval model incorporating hypertext links', in *Hypertext '89 Proc.*, pp. 213–224, Pittsburgh, Pennsylvania (1989).

- 
9. D. Lucarella, 'A model for hypertext-based information retrieval', in *Hypertext: concepts, systems and applications*, eds., N. Streitz, A. Rizk, and J. André, 81–94, Cambridge University Press, Cambridge (1990).
  10. J. M. Smith and D. C. P. Smith, 'Database abstractions: Aggregation and generalisation', *ACM Transactions on Database Systems*, **2**(2), 105–133 (1977).
  11. J. M. Smith and D. C. P. Smith, 'Database abstractions: Aggregation', *Communications of the ACM*, **20**(6), 405–413 (1977).
  12. P. K. Garg, 'Abstraction mechanisms in hypertext', *Communications of the ACM*, **31**(7), 862–870, 879 (1988).
  13. M. Agosti, G. Gradenigo, and P. G. Marchetti, 'A hypertext environment for interacting with large textual databases', *Information Processing and Management*, **28**(3) (1992). To appear.
  14. M. Agosti, A. Archi, R. Colotti, R. M. Di Giorgi, G. Gradenigo, B. Inghirami, P. Mattiello, R. Nannucci, and M. Ragona, 'New perspectives in information retrieval techniques: a hypertext prototype in environmental law', in *Online Information 89*, pp. 483–494, London, England (1989).
  15. R. Colotti, 'Hyperlaw: Prototipo ipertestuale in ambito giuridico', *Informatica Oggi*, **71**(4), 68–73 (1991).
  16. R. Colotti, 'HyperLaw: Prototipo ipertestuale in ambito giuridico', *Informatica Oggi*, **72**(5), 67–72 (1991).
  17. J. Nielsen, 'The art of navigating through hypertext', *Communications of the ACM*, **33**(3), 296–310 (1990).

