Automatic building up documents taxonomy through metadata analysis

Maurizio Talamo, University of Rome "Tor Vergata", Dept. of Mathematics, Rome, Italy
Giorgio Gambosi, University of Rome "Tor Vergata", Dept. of Mathematics, Rome, Italy
Alessandra Aversa, Nestor Lab - University of Rome "Tor Vergata", Rome, Italy
Simone Bonazzoli, University of Rome "Tor Vergata", Dept. of Mathematics, Rome, Italy

Abstract

In cooperation with CNIPA (the Italian Authority for the use of ICT's in the Public Administration), we studied and developed a new solution for the effective access to legal data, especially law texts, norms and rules. Such information represented in XML based and structured documents - is available also at the section or paragraph level. We are experiencing this kind of system within the civil data status, because a project of vertical research, structured on a semantic level, allows the collection of information and the building of a body of uniform rules. The system is based on a statistical similarities relationship and it gives to user the capability to consider also information which, even if not immediately returned as a result of the query resolution, could however be interesting related to the user information needs, because it discovers new information and relationships with in the set of documents. The system provides the usual functionalities of ad hoc retrieval of laws, sections and paragraphs of interest, implemented by means of XML-retrieval techniques, but it also, given the text of a certain law, applies document similarity algorithms to derive section or paragraph, the set of paragraphs where the sections and laws are included which, probably, treat the same subject. Furthermore, by performing a suitable text parsing, the system extracts from each document all explicit references to different laws (and even the references to sections and paragraphs). In this way the system is able, in response to a given query, to return not only all laws (and the corresponding sections and paragraphs) which may be relevant to the specified subject, but also, for each returned law, a set of laws (sections, paragraphs) which are either explicitly (by means of explicit reference in the text) or implicitly (by statistical similarity) related to it. Then these items are ranked by applying a suitable, user tunable, function of both explicit (in a link analysis style) and implicit referent. Applying iteratively the same approach to each considered law, section or paragraph, the user is able to browse within the given document corpus, moving according to the presence of significant (explicit or implicit) relationships among text items. This search technology employs a new class of database designed for exploring information, not just managing transactions, but it lets users prioritize and personalize their choices, rather than directing them down a classification path. Now users can find what they are looking for, and discover new information and relationships. .

Introduction

The modern archiving arrived to the folowing result: the only way to keep the information assets of an archive in its entirety is to keep intact the original structure [8], eg. the creation order of documents during all activities performed by the manufacturer. Therefore, the archive should maintain the "historical method", represented, in Italy, by the implementation of the principle of respecting the original structure of the archives. From the second half of nineteenth century, it went affirming itself in all nations, because it is considered the "[...] most perfect method, indeed the only (the "archivistic method" par excellence for Cencetti[1]) to collate an archive [7]". So the correctness of the information, textual and metatextual, follows from the vision of the documents taken in accordance with the logical sequence of their production. "The sequence of document production that occurs during the activity assigns to each document a prefixed and invariable location in the informative chain of the specific activity"[2] [8], so the alteration of the document place within the sequence inevitably alters the information chain.

The production creates logical and chronological relations among documents which are created by "metatextual information" [3] [8]. The unalterable link among documents in their sequence was defined by Giorgio Cencetti "archivistic link" or "documentary link", understood as "predetermined, original and necessary relationship among documents, in which the documents become evidence of an organic whole, precisely the archive [...]" [8], stressing the lack of autonomy of the archival document individually designed, because it usually has no means when it is separated from previous and subsequent documents and detached from the body to which it belonged [6].

"The need to highlight and preserve the relationships among papers, and research and ensure the easy retrieval of information, encourages the formation of the files, which are the units containing archival documents relating to a specific administrative procedure [...]" [9].

Electronic Document Management System

In the electronic document management it's necessary to study structures showing the conceptual model of the creation and evolution, but focusing the fundamental aspect of maintaining the

[1] GIORGIO CENCETTI (1908-1970), Italian philologist mainly known because his paleography studies, but also because his valuable contributions to the archival discipline during his career as an archivist and professor.

[2] "La sequenza di produzione dei documenti che si verifica nel corso dell'attività assegnata a ciascun documento un posto prefissato ed invariabile nella catena informativa dell'attività stessa".

[3] "dai quali nascono le informazioni metatestuali, quelle, cioè, che pur non essendo materialmente incorporate nel testo dei documenti, si ricavano dal loro confronto ed esame congiunto e la cui correttezza nasce appunto dal rispetto delle relazioni tra i documenti stessi."

history of documents and its link with the earlier and later in the folder. The problem of representing such structures is nowadays more debated in the world. The concept of "archivistic link" is the key to this problem: here, we describe this issue related to the problems faced by computer science and modern archives. Starting from the premise that the core of a computer science research should be to worry about preserving the archivistic link among archival documents managed in a system for documentary information, we will try to establish a valid methodology research, satisfying the different needs but, above all, respectful of this fundamental and mandatory requirement.
In an electronic document management system the context is established in a specific virtual function by the software: the data produced will automatically create a virtual folder, these data are defined "metadata". In a digital environment their importance is significant, because they provide information on the context and structure of a document, they are essential to make it understandable and usable and they must always be kept together with the document they refer to.
Metadata can be divided into three main categories:

- **descriptive metadata**, used for identification and retrieval of digital objects, consisting of descriptions of the source documents, or documents born digitally. This document generally reside in the databases of IR (Information Retrieval) systems;
- **administrative and management metadata**, showing the way to store and maintain the digital objects in system. In the digital world, given the liability of electronic information, these types of metadata can record the technical processes associated with the permanent preservation, providing information on the conditions and access rights to digital objects, certifying the authenticity and integrity of content, recording the custody chain of objects, identifying them clearly;
- **structural metadata**, connects the various components of resources for adequate and full fruition. These metadata also provide data for the documents identification and location, such as the identification code, the address of the files on the server, the proper digital archive and its address.

For a digital archive preservation is needed, in addition to a sensible training of documents at the time, the keeping of the information source of the documents and metadata description, because their specific functions are several.
We currently do not have a software system can including a search request in natural language that find the information requested, accompanied by all of those references that allow you to recreate the archivistic link of the document containing the information. The query must be, first, "translated" into a query processed by the system that, only in this way, can find the object of research.
The scenario that we want to consider is a community of individuals who have a set of skills, eg. a scientific research community, that are used to develop a shared knowledge base. In such a context it is necessary to manage the community's shared knowledge: this knowledge is usually a set of information represented by documents and, as we have seen above, the use of metadata for maintaining the structural coherence of the knowledge base is crucial.
In the next section we show the state of art about organizing documents in digital archives. We analyze the classic hierarchical approach and we highlight some weakness of this model like low flexibility and scalability. We show other approaches to these problems: hybrid or bottom-up strategies like statistical learning, ontologies and folksonomies.

State of the art and open problems

The goal of a retrieval information system (IR system) is to find precisely the information most relevant to the user as a result of the query entered. The focus on the relevance of mere information implies, however, that information can also be found detached from the document that contains it and, therefore, may not be accompanied by the documentary data on the context of origin. In traditional IR systems, to recovery information, usually we can adopt the indices that, strictly speaking, are keywords that appear within a document of an archive or part of it, but the use of these keywords reduces the possibility that the system can also retrieve the context and, consequently, users are often forced to reformulate the query several times in order to get what they really want, because the system will normally find the atomic satisfying the query information and not the document with all the links associated with it.
Computer science tried to overcome this problem using the approach based on identifying relations among concepts, which suggested the idea that the original query should be considered as a point of departure from which will follow, in automatic or semi-automatic mode, one or more calls within the system can achieve greater accuracy in the recovery of the information or documents required. This approach is called "query expansion".
Most of these techniques are based on a thesaurus: in computer science we refers to the thesaurus for all the keywords that give you access to a database or to vocabularies - with lists of synonyms - associated to word processing programs. An interesting type of thesaurus is the similarity thesaurus, which models a type of relationship that can be conceptually defined as "proximity", meaning the relationship between two words.
The information obtained in response is decontextualisated, therefore, from the reference document and other documents which are linked to. These system don't recognize the relationship among retrieved documents and the archive isn't able to ensure their identity. To make it easier to understand we take the example of a search carried out with an engine based on conceptual and proximity type, as the subject of our query, the name "Astorina". The result will be a list of answers that contain the name requested, including those that refer to the Milanese publishing house called "Astorina", but also those relating to a fire company and a studio accountant with the same name.
In case we use an active IR system with query expansion, without a doubt we would have a better degree of precision: a greater sensitivity to the meaning of search and the consideration of the word of question, the consideration on the basis of an explicit or implicit feedback, the view of the information on the basis of the source (eg. authority/reliability of sources). However, the result will always be a list of information decontextualised containing the name "Astorina" but without any archive reference . The IR system must be able to move from the simple search to a search of the document context in the folder of belonging, including recovery means that the metadata identifying the document itself and the relationships with other documents (metadata context). The

set of relations defines the border between content quality, difficult to define, and it accurately defines its content through the archivistic link.
So before we consider search engines based on the proximity of the conceptual thesaurus - take that kind of heuristic techniques - is essential to identify the body of systematised and formalized knowledge, which describe in a way that allows a formal approach. The conceptual problem of proximity and the use of a thesauri can be addressed only by understanding the complexity and indissolubility of specific relationships which, when dissolved, wipe out the research.
For administrating a digital archive that will enable the management and updating of classifications we can use two different approaches [10]:

- **hierarchical/enumerative**: a *top-down* taxonomic scheme, where the information is broken down into more specific categories. An approach in which the achievement of each digital object classified requires the use of a single path. This model assumes a universal knowledge and loses meaning in a specific context, because the designer should know all possible classification schemes and his role would be similar to a demiurge.
- **analytical/synthetical**: this approach splits the object into individual concepts (analytical) and provides the rules for using these concepts in the construction of more complex objects (synthetic). In a such context it is possible to develop new classifications and relations.

The first approach provides the definition of a set of metadata during the design phase of the archive (*ex-ante*) which determines the relations of the knowledge base, such as membership of a particular folder; this approach follows a *top-down* paradigm that defines a hierarchical classification of contents. This approach is inflexible, highly centralized and not scalable, so it is rigid and static and it makes difficult managing and updating the classification procedures. Therefore it's necessary to define an alternative model that is able to easily adapt to changes and evolutions of the archives, but at the same time mantains the requirements of their structural consistency, in particular the preservation of relations among documents.
The second approach exploits the users knowledge and experience to classify and organize documents and information. It's possible to address the problem of classification as a methodology which operates on a model of analysis/synthesis. In this scenario, the process of creation and refinement of structural and descriptive metadata follows clearly a *bottom-up* process, so individual users operate on documents metadata in order to improve their accessibility. An emerging *bottom-up* approach for the classification of documents is represented by *folksonomies*[4]. The collaborative activity of users is the foundation of this model that is based on *folksonomies* where users and creators of content themselves provide the classification of documents. *Flickr*[5] and *Del.icio.us*[6] are the most obvious examples of collaborative classification activities and have shown and still show all their effectiveness, the benefits of this methodology are the high flexibility, dynamicity and scalability in the management of data classification and, consequently, in the management of metadata (descriptive and structural).
As noted in [10] hierarchical/enumerative methodologies are effective and efficient in a highly specialized environment and need a great effort (*ex-ante*) for the definition of descriptive and structural metadata. When the amount of data to store and manage grows considerably the administration of metadata become very difficult, even impossible. A *bottom-up* model solves the problem of scalability but introduce other problems, in particular relating to the structural coherence of the archive. Indeed in the case of *Flickr* and *Del.icio.us*, are primarily treated descriptive metadata (useful to search the contents), if the same model was applied to the structural metadata, necessary for the maintenance of the relations among documents, the " anarchic" action of the users could create structural inconsistencies of the archive. To avoid this problem we can define some tools of analysis that the system makes available to propose classifications that are compatible with the digital object investigated and permit its placement within the structure allowing, therefore, a reachability through logical constraints.

Automatic building up documents taxonomy strategies

Classical approaches based on *top-down* methodologies have several problems: low flexibility and low scalability. On the other hand, purely *bottom-up* strategies, like folksonomies, introduce structural coherence problem on the archive, eg. the conservation of the archivist link.
The definition of alternative strategies that grant to overcome problems of the classical model is crucial. First we have to introduce some form of flexibility in the hierarchical model analyzing content of archived documents and allowing the refinement of descriptive metadata. In this regard, in order to satisfy the emerging needs within the scientific community, it is possible to consider the use of systems based on the statistical analysis of the submitted queries executed on the digital archive by the users [1]. This system is known as "statistical learning". This model allows to extend and specialize the existing taxonomies based on real needs, creating new and different classification strategies. In fact, this approach provides the implementation of an inference engine based on statistical similarity relationship [2] and gives to the user the ability to take into account information which, even if not immediately part of the results of the query, can still be interesting compared with the same needs, because this search engine research and build new relationships among information and documents in the digital archive.
The latest trends in computer science for describing the contents of digital objects look toward those who are called ontologies, which are a formal representation of a set of concepts within a domain and the relationship among the concepts themselves. An ontology is somewhat similar to a *thesaurus*, except that the former is hierarchical in two or more levels, where the top level defines the context, the latter is just a unstructured collection of words. Statistical learning allows to build ontologies that specializes existing classifications and building new relationships.
Our systems provides [1, 2] the usual functionalities of ad hoc retrieval of laws, sections and paragraphs of interest, implemented

[4] It is the fusion of the words *folks* and *taxonomy* and it means that the classification and management (*taxonomy*) is performed by the common people (*folks*).
[5] http://www.flickr.com
[6] http://www.del.icio.us

by means of XML-retrieval techniques, but it also, given the text of a certain law, applies document similarity algorithms to derive section or paragraph and the set of paragraphs where the sections and laws are included which, probably, treat the same subject. In this way the system is able, in response to a given query, to return not only all laws (and the corresponding sections and paragraphs) which may be relevant for the specified subject, but also, for each returned law, a set of laws (sections, paragraphs) which are either explicitly (by means of explicit reference in the text) or implicitly (by statistical similarity) related to it. This kind of research method using a new class of database that allows end users to customize and assign a priority to their needs, without, however, having to act directly on the classification and preservation of relations among documents.

When we introduce some degree of flexibility to allow the definition of structural relationships among documents what problems could we encounter? The *top-down* approach doesn't permit the dynamic modification of structural metadata, while the *bottom-up* approach permits that dynamicity but introduces the archive coherence problem described before. Therefore it's necessary to adopt an approach that uses folksonomies benefits, quality of the statistical learning approach and, at the same time, could be able to maintain the archivistic link.

Conclusions

In this work we explain several approach to the solution of electronic document management problem in a digital world: the hierarchical model isn't flexible and the organizer and designer must have a universal knowledge about documents and its relationships. The folksonomies approach efficiently uses scientific community knowledge and skills; such approach works quite well when it's used to describe and classify documents content (*Flickr* and *Del.icio.us* and, in general, search engines), but it need to be addressed by predetermined logic to build a consistent and coherent organization and classification model.

It is therefore necessary to "guide" the process of creating and managing metadata in order to prevent this kind of problem. A possible solution could be a cooperative, flexible and adaptive model that, starting from the classification of the metadata of a body for the maintenance of relations among documents, allows the evolution of the system through the contribution of those who need to enrich the system by providing information on the navigability and the classification and reporting any inconsistencies that a user could create [11].

An alternative approach could be a model based on an engine that can respect the logical rules to assure a set of coherent structural relationships [11]. Such approach, called *Constraint Network*, is defined by a minimal structural metadata set and it allows to specialize existing classifications keeping structural coherence. The classification evolutions result is an XML schema that comply with semantic links and constructions rules identifying all allowed path to retrieve documents.

References

[1] Giambattista Amati, Edgardo Ambrosi, Marco Bianchi, Carlo Gaibisso, Giorgio Gambosi, Automatic Construction of an Opinion-Term Vocabulary for Ad Hoc Retrieval, ECIR, pp.89-100, (2008).

[2] Edgardo Ambrosi, Marco Bianchi, Carlo Gaibisso, Giorgio Gambosi, A Description Logic based Grid Inferential Monitoring and Discovery Framework, GCA, pp.18-23, (2005).

[3] Franco Arcieri, Fabio Fiovaranti, Roberto Giaccio, Enrico Nardelli, Maurizio Talamo, Certifying performance of cooperative services in a digital government framework, in Symposium on Applications and the Internet Proceedings, IEEE Computer Society Press, pp.249-256, (2003).

[4] Franco Arcieri, Elettra Cappadozzi, Paolo Naggar, Enrico Nardelli, Maurizio Talamo, Choerence maintainance in cooperative information system: the Access Keys Warehouse approach, in International Journal of Cooperative Information Systems, 11(1-2), pp.175-200, (2002).

[5] Franco Arcieri, Elettra Cappadozzi, Paolo Naggar, Enrico Nardelli, Maurizio Talamo, Access Keys Warehouse: a new approach to the development of cooperative information systems, in 4^{th} International Conference on Cooperative Information Systems Proceedings, IEEE Computer Society Press, pp.46-56, (1999).

[6] Giorgio Cencetti, Scritti Archivistici, Il centro di ricerca editore, p.64, (1970).

[7] Elio Lodolini, Archivistica. Principi e problemi, Franco Angeli, p.213 (2005).

[8] Luigi Londei, Elementi di archivistica, Jouvence, (2003).

[9] Stefano Pigliapoco, La memoria digitale delle amministrazioni pubbliche, Maggioli Editori, (2005).

[10] Emanuele Quintarelli, Folksonomies: power to people, ISKO Italy-UniMIB meeting, (2005).

[11] Maurizio Talamo, Simone Bonazzoli, Constrain Network: a tool for coherent building up documents taxonomy, to be published, (2009).

Author Biography

Maurizio Talamo, full Professor of University of Rome "Tor Vergata", Faculty of Mathematical, Physical and Natural Sciences; President Nestor Lab and "INUIT-Tor Vergata" Fondation, author of registered patents and original models published on authoritative international reviews; member of many governative boards.