Knowledge Mining and Discovery for Searching in Literary Texts

A. Cappelli* +, **M. N. Catarsi***, **P. Michelassi***, **L. Moretti***,
*Istituto di Linguistica Computazionale – CNR Pisa
M. Baglioni+, **F. Turini**+
+ Department of Computer Science, University of Pisa
M. Tavoni°,
° Department of Italian Studies, University of Pisa

Abstract

The article describes a query system on texts and literary material with advanced information retrieval tools suitable to retrieve the content of a text, either as material specifically organized with respect to linguistic, stylistic and rethoric features, and in its historical, social and cultural context. As a test bed we chose the Dante's characters of *al di là*. This method of investigation should help a scholar of a literary text to realize part of his interpretative intentions. For this purpose, we will adopt advanced methodologies in knowledge management and knowledge discovery to be applied to a rich representation of the tagged content of a text.

Introduction

One of the key conditions for transforming large quantities of texts into effective repositories of electronic information is the ability to "search by an idea", as opposed to the more or less sophisticated pattern (i.e. strings or keywords) matching capabilities built inside most of the current available IR systems (European Community, 1992). Conceptual search is strictly related to the problem of "active content" which, from a technological viewpoint, consists in the development of techniques and tools to effectively access and use the content of the huge material digitally stored in global information networks. In this perspective, it is important to design tools for the creation, analysis and use of written texts in such a way as to leverage linguistic knowledge embedded in texts for creating, in synergy with other relevant media, an organization of the information modelled for the creation and transmission of knowledge.
The quality of the tools suitable for information retrieval thus becomes a strategic factor to give the possibility to access the huge mass of increasing information stored in global networks in an effective way by using information retrieval techniques based on complex conceptual processes. This problem is currently relevant in the information society, since the possibility of accessing knowledge can enable any human operator, in any part of the world, to react, in real time and effectively, to the economic and cultural market. It is well known that the success of many human initiatives in different domains, strictly depends on the possibility of acquairing adequate information about the economic, social and cultural context on which to act.
Our work aims approaching the cultural phenomenon through the development of new methods and more power tools for the retrieving of the content of digitally stored literary material. This activity can also produce benefits in the field of cultural heritage and, in particular, in the production of electronic tools to be used by a large variety of subjects, from scholars to students, to those generally interested in cultural products for several aims, either for didactic, entertainment or advanced scientific investigations.

The aim of our project is to build a query system on texts and literary material with advanced information retrieval tools suitable to retrieve the content of a text, either as material specifically organized with respect to linguistic, stylistic and rethoric features, and in its historical, social and cultural context.
As a test bed we chose the Dante's characters of *al di là*, a domain consisting in a set of data and relations so complex to be adequate to match the above requirements. It is well known among Dante's scholars that the "interpretation" of the Commedia requires the evaluation of many types of knowledge regarding both the structure of the text and the general context in which it has been generated. The context has to be intended as the set of historical, political, social, philosophical and cultural experiences of Dante which have been explicitly or implicitly "used" in the production of the Commedia.
This method of investigation should help a scholar of a literary text to realize part of his interpretative intentions. For this purpose, we will adopt advanced methodologies in knowledge management and knowledge discovery to be applied to a rich representation of the textual content.

System specifications

A system able to analyze and to interpret a text in accordance to the above mentioned assumptions, must adapt itself to a scholar behaviour in expressing informative intentions about many aspects of a text. Consequently, this system must exhibit an expressive power able to analyze and to understand complex queries a scholar may formulate about a domain.
The system should then be able to provide the scholar with useful answers. That is answers not so deep as the ones given by a scholar with all the data at his disposal, but at least computed with respect to all available data, that can not be held all together in the memory of the scholar; we also expect that the answers are not simply present in Dante's encyclopedia.
The system must be able to answer, for example, to the following queries:

- How many and who are the *ghibellini* and how many and who are the *guelfi* in each single partition of the Purgatory and Paradise, and in each of the *cantiche* all together?
- Which are Dante's attitudes towards holders of feudal power, in a decreasing order?
- Point out, by using different colours, on a Europe map the clusters of characters in *Inferno*, *Purgatorio* and *Paradiso*;
- How many and who are the mythological characters that are referred to in *Inferno, Purgatorio* and *Paradiso*;
- Is there a statistical correlation between the belonging to a feudal system and salvation, or between belonging to city system and damnation;
- A sorted set of links in the text, dealing with the noble families from *Romagna*.
- Dante is more kin with religious or secular clergy.
- Which are, in the two centuries before Dante, the years mostly characterized by corruption, which are the cities and which are the institutions. Which are the years and the cities in which the fights among parties have been more bloody.
- The Greek and Latin poets and the vulgar poets met by Dante in the *al di là* speak all the same way, or they have different lexicon and syntax;
- Dante's judgment about an institution is coherent with his judgment about the members of the institution.

Answering these queries requires the evaluation of different types of knowledge. In general, while some of them can be solved by the evaluation of the knowledge of the content of the text, others require a comparison between the content of the text and that of the context. Furthermore, some answers are computed by evaluating only the context. In other words, answering these queries means computing complex and sometimes sophisticated processes applied to the conceptual aspects of Dante's world.

A computational system able to implement such conceptual processes must be provided with these kinds of knowledge and must be able to dynamically manage them in accordance to the specific goals expressed in a query.

The computer in literary studies

So far, the use of intelligent techniques in humanistic fields and, in particular, for the historical and literary analysis, did not receive a particular attention by researchers.

The use of the computer in literary studies has been in general devoted to the production of low structured material (concordances) to aid a scholar to formulate general hypothesis or induce models about the lexical organization of a text via a non automatic process. Recently, with the design of new tools for text representation, new perspectives have been raised which allow more expressive operations on text. For instance, by structuring text with advanced markup or hypertext tools, certain conceptual manipulations of a text become more and more possible.

The markup of several aspects of a text, starting from typographical conventions to more abstract semantic or interpretative features, is the goal of the Text Encoding Initiative. By using SGML as a high-level markup language, texts can be stored thus obtaining a rich metarepresentation of their multilevel information. (Sperberg-MaQueen & Burnard, 1994)

Hypertext representation can be used like a "high-level declarative programming language" for the realization of a theory of narrative evolution, typical of certain literary trends (Sutherland, 1990). In this sense, a representational formalism "promise to expose and even deal with a text's excessive textuality or with what lies beyond its textual confines" (ibidem). In this sense, textual information can be profitably linked with several "interpretative" levels lying beyond the text, which can cope with the goals of a critic's attitude that aims at relating textual and contextual information.

In "The World of Dante" Project at the Virginia University, a hypermedia environment for the study of *Inferno* of *Divina Commedia* has been implemented. The text has been tagged by using SGML and a rich data base of many aspects of the world of Dante is availabe which enables a scholar to formulate answers about characters, events, etc.

New approaches to literary computing by using new computational means suggest the possibility for the computer to perform "not simply as a concordancing and quantifying program though which text is fed, but as an interactive environment adaptable to the constant desire of literary critics to inhabit and explore a cultural and aesthetic space" (Sutherland, 1990).

Computer programs based on propositional calculus have been implemented in order to interpret several semantic aspects of a text, such as, for instance, methaphor, whose meaning has been computationally specified (Steinhart, 1995).

A new perspective is raising towards a computational interpretation of a text based on the treatment and understanding of its meaning. This in turn requires the treatment of the knowledge "not only of the text itself, but of the world" (De Vuyst, 1990).

New formalisms for representing the content of a text and those for the representation and the discovery of knowledge are thus useful in this perspective.

A rich metarepresentation of a text produces a storage of all its multidimensional aspects, and allows one to analyse eiher its specific linguistic, stylistic or rethoric organization or its general cultural context.

This is also the base for the specification of advanced knowledge management tools for the treatment of textual content. Traditional text processing procedures, such as, for instance, text indexing, can thus be converted in sophistacated knowledge management processes.

Conceptual search

Current information retrieval systems exhibit many limits, even if they use linguistic parameters to control searching processes. In particular, it is not possible to find the information needed because words or phrases used in queries are sometimes different from those used to store the material to be searched for. Furthermore, the indexing of a text on the basis of strings of characters which, in a more or less sophisticated manner characterises most part of current commercial systems, does not allow to search information on the basis of conceptual methods involving

complex linguistic and conceptual competences, both generic and specific. While it is possible to search clusters of words on the basis of semantic processes involving lexical relations (hyponymy, hyperonymy, characteristics, etc.), it is quite impossible to formulate queries involving the knowledge of the contents of a text and of its domain.
As to searching into literary texts, while it is necessary to retrieve entire documents or parts of them on the basis of linguistic, lexical or grammatical parameters, it would be also suitable to search on the basis of the inferential processes involving the general knowledge of an author, an historical period, a literary style and also of the relationship between this general information and the linguistic and stylistic realizations of a text.
The realisation of this "searching by an idea" method, requires linguistic and knowledge engineering at any levels, from the formulation of the queries to the organization of the documents, both of their form and contents. It also requires a suitable representation of the domain, which consists in the choice of an adequate descriptive terminology structured in an organization so cohesive to account for the nature of concepts and of their mutual interrelations.
These steps are strictly related since it is not possible to effectively express an idea without suitable expressive means, while it is not possible therefore to search for the contents if not adequately represented.
Representation of the contents of a text, representation of the domain and search mechanisms thus become three strategic points in this perspective.
A basic procedure of text processing is indexing. If the content is the target of retrieval, a text has to be indexed in accordance to its content in order to enable the user to use problem solving and good intuition concerning relationships between words and, in particular, about the ways they are used in the indexed material.
Conceptual indexing is the aim of some systems developed at Sun Microsystems, which combine knowledge representation and natural language processing techniques with traditional document indexing tools. The aim of such systems is to enable an information retrieval system to discover a hidden terminology from texts, on the basis of a terminology used by a user in its queries (Woods, 1997).
A prototype of a conceptual information retrieval system has been implemented by the Project LRE CRISTAL (Conceptual Retrieval of Information using a Semantic dicTionary for Access in three Languages) which allows to retrieve documents about economic press in French, by using Italian, French and English to express queries (Cappelli, 1994, 1997). Documents are indexed on the basis of word meanings and queries allow to search in accordance to such criterion. The disambiguation of the words in the documents and in the queries is performed by the exploitation of a conceptual dictionary and by the calculation of distance metrics between word meanings present in specific chosen contexts (Magadur & Tabuteau, 1996).
Many sytems make use of advanced linguistic tools, such as, for instance conceptual dictionaries, for enhancing and refining queries (Croft & Smith, 1992; Voorhees, 1993, 1994; Turtle & Croft, 1991; Qiu & Frei, 1993; Paice, 1991; Kristensen, 1993; Jones et al., 1995; Järvelin et al., 1996; Hancock-Beaulieu, 1992; Croft, 1987; Chen et al., 1993). A representation of the lexicon in a complex conceptual map allows to implement deep procedures for the treatment of the semantic aspect of a text (Miller et al., 1990).
Many experimental information retrieval systems use knowledge based techniques developed in the framework of artificial intelligence. Such systems are applied in several fields, such as, for instance, in medical diagnosis, in engineering systems faults, in decision making, etc. (Hayes-Roth & Jacobstein, 1994; Monarch & Carbonell, 1987; Vickery & Brooks, 1987; Smith et al., 1989; Cohen & Kjeldsen, 1987; Fox, 1987; Pollitt, 1987; Chiaramella & Defude, 1987; Chen & Dhar, 1991).

Data mining and knowledge discovery

Recent approaches try to investigate and use knowledge mining and discovery techniques. Knowledge discovery is the process of extracting useful and novel knowledge from large databases (Fayyad et al., 1996; Piatetsky-Shapiro.& Frawley, 1991; Han & Mamber, 2000). The key step in this process is data mining, which consists in the application of several techniques for inducing models (or profiles or patterns) from data. Such models are used for highlighting regularities, characteristics, rules and associations hidden in data. Text mining is a new research area that tries to solve the information overload problem by using techniques from datamining. Text mining involves the preprocessing of document collections (text categorisation, terms extraction), the storage of intermediate representations (distribution analysis, clustering, trend analysis, association rules etc.) and visualisation of results. Text mining is aimed at extracting a large number of features that represent each of the documents. Its peculiarity is the need for background knowledge; in fact even patterns supported by a small set of document may be significant. For instance, key-words based association analysis collects sets of keywords or terms which occur frequently together and then finds the association or correlation relationships among them. Automated document classification is an important text mining task where first, a set of preclassified documents is used to derive a classification schema in the form of decision tree. Such schema is then used for classification of other on-line documents.

Document representation

A very important task of the approach we are describing concerns the individuation of an adequate representation of the content of a text. For certain realizations we discussed above, the indexing of a text on a conceptual basis is the key.
Markup languages constitute good tools to declaratively represent textual data with their mutual relationships in complex data structures progressively and dinamically "evolving". Once tagged, a text becomes a rich inductive base for computer manipulation.
XML (eXtensible Markup Language) is a SGML (Standard Generalised Markup Language) subset. It is an emerging standard for the representation and exchange of documents, since it makes it possible to associate some

information to the documents regarding, for instance, the their structure, components and the relations between them. This class of languages is suitable to account for dynamically structuring a literary text, whose interpretation is dependent from specific individual needs. A markup language has in fact no predefined tags but works on a general metalanguage on the basis of which new language can be defined. The "a priori" fixed XML tags are limited to the control information and the user can arbitrarily define his own tags: this can be done using a formal model known as Document Type Definition (DTD). This feature makes XML a metalalnguage with which is possible to define "ad hoc" markup languages for applications with specific objectives. In XML does not exist a predifined tag set and so does not exists a predefined semantics for tags, but the semantics is defined by the application that processes the XML document. This characteristic allow the exchange of data between components of a distributed system: it is sufficient that the system components share the same DTD to elaborate documents and access to the document structure.
This extensibility will allow to dynamically increasing textual metarepresentation by adding new tagged information according to new emerging aims.
On a XML database, suitable search mechanisms can be realized which retrieve information on the basis of a parametric specification of its structure. More advanced mechanisms can also be obtained by an integration of such procedures with data mining tools through which, with the exploitation of different types of background knowledge, an interpretation of the XML database can be achieved.
A query answering process can be specified in which both invocations to mining tools and the use of domain knowledge can be integrated like in more traditional fields of Knowledge Discovery in Data Bases. As a result, an XML based environnment can be designed in which several data mining tools can interoperate and pre-defined domain knowledge is represented.

Architecture

Given all aspects described so far, the system we propose behaves in the following way. The query processing component interacts with the knowledge base and retrieves existing knowledge or it creates novel information from it.
The architecture of the system is composed by the following modules (see figure 1):

- **User Interface** for formulating queries of the types specified in the specifications of the system and to visualize their results;
- **Query Manager** which evaluates the queries in accordance to their specific goals;
- **XML Query Executor** to execute queries which do not need the firing of data mining procedures;
- **Data Mining Executor** to execute queries which need the firing of data mining procedures;
- **Repository** to store the results of the queries;
- **Knowledge Base** composed of the two following partitions:
  - **Meta-Text** for the metarepresentation of a text in XML format;
  - **Ontology** in which domain knowledge is organized;

The role of the interface is to create chunks of conceptual elements to be used by the search mechanism which, in turns, will map conceptual clusters, produced by the analysis of a query, onto the two partitions of the knowldege base, in accordance to different informative goals expressed in the query. In the search module, retrieval, data mining and discovery tools are triggered they exploit background knowledge of the domain and knowledge procedurally extracted from the metarepresentation.
The knowledge base contains the concepts, which adequately describes the domain and the metarepresentation of the text.

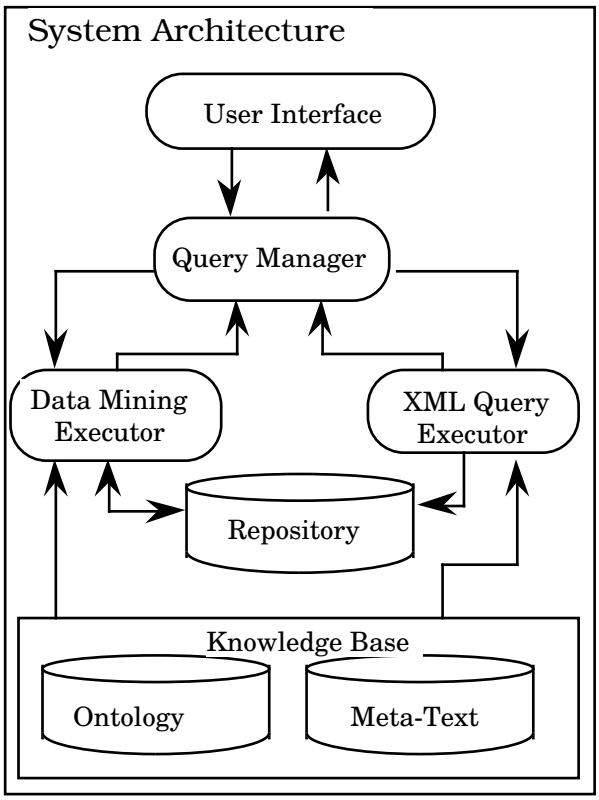

Figure 1: System Architecture

Structure of the Knowledge Base

The knowledge base has been structured taking into account the general partition between textual and contextual knowledge, we introduced to realize our "interpretation hypothesis".
Furthermore, other distinctions can be established in accordance to the different modalities by which knowledge can be specified. A distinction between a declarative and a derivative part of knowledge, both textual and contextual, can be drawn following the way by which knowledge is inserted into the knowledge base: declarative knowledge is specified in an "a priori" way, while derivative knowledge is that inferred by applying search or mining mechanisms.

Metarepresentation of the text

The text of Divina Commedia is represented in XML format.
Our final goal is to build, via a systematic and exhaustive XML mark-up of Dante's Divine Comedy, a meta-text containing a large amount of semi-structured data concerning the characters of *al di là*: that is all the people, places and institutions dealt with or even only named in the text. This meta-text will be used to feed the knowledge base that, via the connected query processing process, will allow the interpretation of user questions and the construction of intelligent answers.
Anyway, "all markup of written texts faces the choice between focusing on uninterpreted transcription of layout and typography, and interpretive translation of a printed representation into an electronic representation", but it is this interpretive translation which only provides means to facilitate computer processing directed to specific adavanced ends other than simple display (Lavagnino & Mylonas, 1995).
In order to avoid the realization of a text metarepresentation useful only for very specific and too narrow goals, a number of principles for individuating and organizing data have been chosen.
Among the many types and combinations of data that might have been chosen, we selected types of data and relations satisfying the following requirements:
- Being objective and easy enough to be extracted from specific points in the text;
- Being in a finite number: a number large enough to offer a good amount and variety in contents, but not so large to be unmanageable;
- Forming a collection of encyclopedic knowledge of the text that is useful and significant even if only simple text retrieval tools are available;
- Including information on the items and relations that are not simply facts mechanically verifiable; this means that the information has to be acquired from competent scholars;
- Allowing an effective transfer of knowledge from the expert to the system, in the form of elements of evaluation and judgement, that the system is required to categorize and interrelate.

Starting from these general principles, certain specific criteria can be derived in order to define a proper XML mark-up of the text to be linked to the organization of the domain knowledge, such as, for example:
- Identify all the characters, the places and the institutions occurring in the text both when they are named and when they are referred to via periphrase;
- Partition the whole text with respect to the topology of the three reigns, and, hence, indirectly with respect to the moral classification of vices and virtues, in such a way that all the characters and the objects are always linkable to the area of text they belong to;
- As to *al di là* characters, point out their sins and their personal attitude;
- Note, for each character the attitude of Dante towards the character and vice-versa (aggressive/sympathetic, scornful/respectful etc);
- Note, for each character, which is its style (in terms of behaviour, rhetoric, ecc.) and which is its way of speaking (in terms of lexicon, syntax etc.);
- Note, for each character, which is the judgement of Virgilio and Beatrice;
- Note, for each character, which are the other characters related in the text, let it be in an explicit way via similes and other rhetoric figures, let it be in an implicit way via simple closeness associations;
- As to characters, partition them in two fundamental categories: characters in the *al di là*, and other characters;

Text	**Metatext**
Nel mezzo del cammin di nostra vita mi ritrovai per una selva oscura che la diritta via era smarrita. Ah quanto a dir qual era è cosa dura esta selva selvaggia e aspra e forte che nel pensier rinova la paura! Tant'è amara che poco è piú morte; ma per trattar del ben ch'io vi trovai, dirò dell'altre cose ch'i' v'ho scorte. [...] Mentre ch'i' rovinava in basso loco, dinanzi alli occhi mi si fu offerto chi per lungo silenzio parea fioco. Quando vidi costui nel gran diserto,	Narrazione Inizio Verso:1 Carattere:1 Fine Verso:64 Carattere:36 Collegamento Dialogo Inizio Verso: 65 Carattere:1 Fine Verso:136 Carattere:37
"Miserere di me" gridai a lui, "qual che tu sii, od ombra od omo certo!"	Segmento ChiParla: Dante AChi: Virgilio Testo: <<Miserere di me qual…>> Inizio: Verso: 65 Carattere:1 Fine Verso: 66 Carattere:43 Interruzione: "gridai a lui" Inizio:… Fine:…
Rispuosemi: "Non omo, omo già fui, e li parenti miei furon lombardi, mantovani per patria ambedui. Nacqui sub Julio, ancor che fosse tardi, e vissi a Roma sotto 'l buono Augusto nel tempo delli dei falsi e bugiardi. Poeta fui, e cantai di quel giusto figliuol d'Anchise che venne di Troia, poi che 'l superbo Iliòn fu combusto. Ma tu perché ritorni a tanta noia? perché non sali il dilettoso monte ch'è principio e cagion di tutta gioia? "	Segmento ChiParla:Virgilio AChi: Dante Introduzione: "Rispuosemi:" Inizio:… Fine:…. Testo: <<Non omo, […] tutta gioia?>> Inizio:… Fine:…
[…]	[…]
"Poeta, io ti richeggio per quello Dio che tu non conoscesti, acciò ch'io fugga questo male e peggio, che tu mi meni là dove or dicesti, sí ch'io veggia la porta di san Pietro e color cui tu fai cotanto mesti". Allor si mosse, e io li tenni dietro.	Segmento Introduzione: E io a lui: Inizio:…. Fine:…… Testo: "<<Poeta […] mesti>>." Inizio: Fine:…… Conclusione: "Allor si mosse […] dietro." Inizio:….. Fine:….

Figure 2: Representation of the text

- As to historical characters, categorize them with respect to periods and environments (middle-age, ancient Greece, ancient Rome), place, chronology, category and sub-category (clergymen>popes>bishops>priests>church people and so on; emperors>kings etc.); as to literature characters, categorize them with respect to the literature they belong to (Bible characters, mythology, romance literature etc.);
- As to characters, both real and from the literarature, provide them with further information like: relationships with places and institutions, relationships of friendship/contrast with other characters;
- Note for each town, country, corporation, dynasty etc. the explicit judgement of Dante.

The above instructions will be partly exploited as formal characterizations (every time they are simply unique objects, like names, dates and so on, representable as a list of terms), and partly they will be represented as descriptions (like the many relations that may link a character to another one, to a polytical party, to a corporation, to a town, to an institution ect.).
The system, this way, will be fed with real knowledge items, that is categorizations of elements and their relations provided by human experts. The above mark-up operations require a high degree of competence in contents and judgments from the scholars and it is currently achieved by hand with the guide of the perspicuous presentation of the meta-structure of the knowledge base.
In figure 2, a schema for the representation of the text in XML format is shown. In particular, part of the representation of the dialogue between Dante and Virgilio is presented. The information which has been tagged concerns, among others: who are the actors of the dialogue, what they say, in which part of the text they act, what are the allocutionary modalities used by a character, what is the development of the dialogue and in which segments is diveded (introduction, interruption, conclusion), etc.

Ontology

In the ontology, all concepts describing the domain knowledge are represented, with their relationships.
The system should structure knowledge in accordance to logical relationships between the objects (e.g. the relationship that connects all dominicans to the dominican order, all the holders of feudal rights to the feudal order, the relationship among them of all the elements belonging to the same class etc.) and classify relationships among characters and objects, by recognizing subsumption relationships among relationships (e.g. hostile relationships, more or less strong, between a character and a town, etc.) Summing up, the system will be able to handle the whole body of information, moving from single elements and single relationships to the whole system of relationships.
Conceptually, any object is part of a is-a hyerarchy through which subsumption principles are valid. Each object is locally described by a set of characteristics in accordance with the general DTD. The values of the characteristics can be either single elements or informal descriptions. The general schema of the representation of the ontology , of the meta-text and of the relationship between them are shown in figure 3.

Query Manager

The text and the ontology are evaluated in a dinamic way but, in any case, following two possible modalities:
- searching through data and returning stored values;
- searching through data and inferring new values.

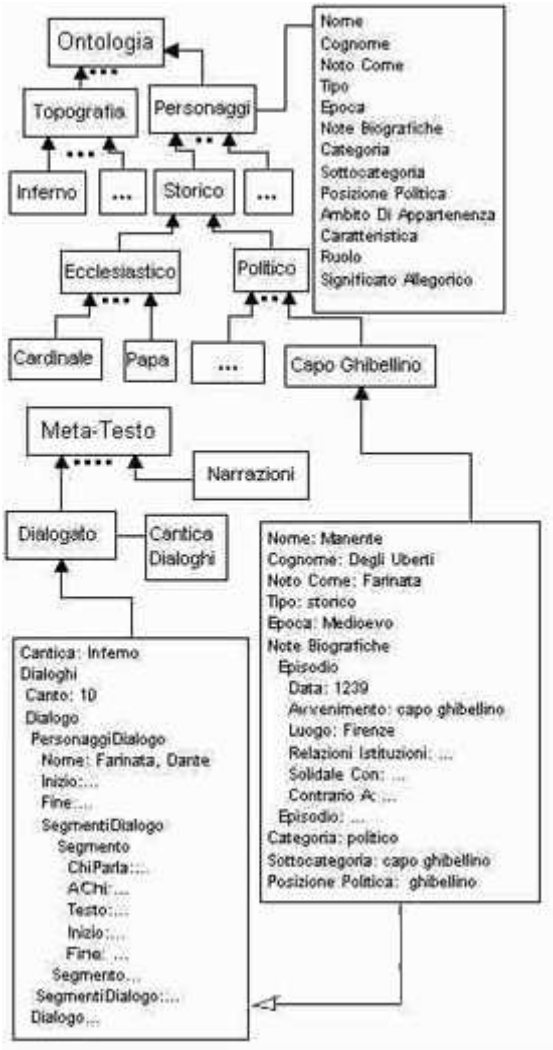

Figure 3: Logical schemas

As to the former modality, tools retrieving XML data are used, which return values of tagged information, both textual or ontological. As to the latter, data mining and discovery tools are applied either to the knowledge base itself or to the results returned by a previous application of XML searching tools.
X-Query and KDDML languages are used for performing, respectively the two above mentioned tasks (Alcamo et al., 2000).
KDDML is an environment based on XML in which is possible to combine several types of knowledge extraction operations.

Let us introduce an example in order to clarify the modality by which queries are computed.
To compute the following query:

Classify religious characters in accordance to the Dante's attitude towards them

a classification tree is needed which will be created in accordance to some qualitative rules explaining degrees of attitude an author can have towards the characters (compession, hatered, likeness, etc.).
So, the processing of this query requires two steps.
In the first, religious characters with the Dante's attitudes towards them are searched into two different areas:
- religious characters are searched into the domain ontology;
- Dante's attitudes are searched into the metarepresentation of the text.
The two tasks are realized by applying XML retrieval tools.
An example of such query is shown in the following.

```
XQUERY

<Table>{
for $p in
(document("Ontologia.xml")//Personaggio)
let $a:= (document("Testo1.xml")//Dialoghi)
where $p/Categoria="ecclesiastico" AND

$p/NotoCome=$a/Dialogo/PersonaggiDialogo/Nome
return
<Religioso>
   {$p/Nome}
   {$p/Cognome}
   {$a/Dialogo/AtteggiamentoDante}
</Religioso>
}
</Table>
```

This query searches through the Ontology and Metatext databases returning religious characters and the Dante's attitudes towards them, as shown in the following query result fragment:

```
<Table>
   <Religioso>
      <Nome>Alberigo</Nome>
      <Cognome>dei Manfredi</Cognome>
       <AtteggiamentoDante>duro, spietato,
sprezzante</AtteggiamentoDante>
   </Religioso>
       É..
   <Religioso>
      <Nome>Piccarda</Nome>
      <Cognome>Donati</Cognome>

<AtteggiamentoDante>rispettoso</AtteggiamentoD
ante>
   </Religioso>
</Table>
```

The second step of the query is performed by the evaluation of a classification tree, in which Dante's attitudes towards religious characters are intepreted. The following is an example of such a query using KDDML language.

```
<?xml version = "1.0" encoding = "UTF -8"?>
<DOCTYPE KDDML_OBJECT SYSTEM "kdd.dtd">
<KDDML_OBJECT>
   <KDD_QUERY
name="ClassificaAtteggiamento.xml">
   <CLASSIFY_TUPLE
xml_dest="Atteggiamento_Classificato.xml">
    <FILE_ARFF file_name="Religiosi.arff"/>
   <FILE_TREE
file_name="Atteggiamento_Albero.xml"/>
   </CLASSIFY_TUPLE>
</KDD_QUERY>
</KDDML_OBJECT>
```

The rules interpret data tagged in the KB in accordance to a chosen model. Their aim is to infer regularities from some heterogeneous occurrrences in the KB. As an example, the fact that Dante smiles, or is kind when he meets a character can both be interpreted as a "positive attitude": an interpretation which lies beyond their different surface realizations.
The result of the interpretation can be added to knowledge of the system, thus increasing the inductive base for the achievement of new interpretive goals.

Conclusions

The application described in this article shows that text processing can be profitably improved by the integration of advanced knowledge representation tools. Even literary texts, whose structure is inherently complex, due to the subtle elaboration of every linguistic level, can be interpreted from a semantic viewpoint, once an adequate representation of their domain is specified. Current knowledge representation languages are able to account for the structure and the relationships of the world of a text and of its "external" context. Moreover, they are able to discovery inherent conceptual chunks hidden in the representation.
The system is then be able to provide a user with useful answers with respect to all data at its disposal.
Answers are tipically computed through the computation of different types of knowledge, each type pertaining to a different partition in the conceptual organization of an author's world.
This has many consequencies on two possible levels of analysis:
- the organization of the mental space a scholar build about an author;
- the specification of the tools able to threat such a space.
As to the former, the ways through wich a scholar organizes a domain with respect to his "subjective" goals have to be furtherly investigated.
As to the latter, while current available tools are able to distinguish between different conceptual partitions, the modalities through which different conceptual areas can be computed, still remain a challenge.

References

Alcamo P., Domenichini F., Turini F. (2000). An XML Based Environmenti in Support Overall KDD Process. In Proceedings FQAS 2000 (pp. 413-424).

Cappelli A. (1997). Accesso Multilingue al Contenuto dell'Informazione. La Comunicazione, Numero Speciale Atti del Convegno, Trattamento Automatico della Lingua nella Società dell'informazione, XLVI (1.2.3.4).

Cappelli A. (1994), CRISTAL Conceptual Retrieval of Information using a Semantic Dictionary for Access in Three Languages. AI*IA Notizie VII (3), 45-50.

Chen H., Dhar V. (1991). Cognitive processes as a basis for intelligent retrieval system design, Information Processing and Management. 27 (5), 405-432.

Chen H., Lynch K. J., Basu K., Dorbin T. (1993). Generating, Integrating, and Activating Thesauri for Concept-Based Document Retrieval. IEEE Expert, Special Series on AI in Text-Based Information Systems, Vol. 8, N. 2, 25-34.

Chiaramella Y., Defude B. (1987). A prototype of an intelligent system for information retrieval: IOTA, Information Processing and Management, 23 (4), 285-303.

Cohen P. R., Kjeldsen, Information retrieval by constrained spreading activation in semantic network, Information Processing and Management, 23 (4) 255-268.

Croft W. B. (1987). Approaches to intelligent IR, Information Processing & Management, Vol. 23, N. 4, 249-254.

Croft W. B., Smith L. A. (1992). A Loosely-Coupled Integration of a Text Retrieval System and an Object-Oriented Database System. In Proceedings of the 15th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval.

De Vuyst, J. (1990). Knowledge Representation for Text Interpretation, Literary and Linguistic Computing, 5 (4), 296-302.

Dutoit D. (1992) A Set-Theoretic Approach to Lexical Semantics. In Proceedings of COLING-92 (pp. 982-987).

European Community. (1992). LRE Thematic Background Documents.

Fox E. A. (1987). Development of the CODER system: A testbed for artificial intelligence methods in information retrieval. Information Processing and Management. 23 (4), 341-366.

Han J., Kamber M. (2000). Data mining: concepts and techniques. Los Altos (Ca.): Morgan Kaufmann.

Hancock-Beaulieu M. (1992). Query Expansion: Advances in Research in Online Catalogues. Journal of Information Science, Vol. 18, 99-103.

Hayes-Roth F., Jacobstein N. (1994). The state of knowledge-based systems, Communication of the ACM, 37 (3), 27-39.

Järvelin K., Kristensen J., Niemi T., Sormunen E., Keskustalo H. (1996). A Deductive Data Model for Query Expansion. In H. P. et al. (Eds.) Proceedings of the 19th Annual International ACM SIGIR Conference on Research and Developement in Information Retrieval (pp. 235-243). Zurigo.

Jones S., Gatford M., Robertson S., Hancock-Beaulieu M., Secker J., Walker S. (1995). Interactive Thesaurus Navigation: Intelligence Rules OK. Journal of the American Society for Information Science, Vol. 46, (1), 52-59.

Kristensen J. (1993). Expanding End-Users' Query Statements for Free Text Searching with a Search-Aid Thesaurus. Information Processing & Management, Vol. 29 (6), 733-744.

Lavagnino, J., Mylonas E. (1995). The Show Must Go on: Problems of Tagging Performance Texts. Computers and the Humanities, 29, 113-121.

Magadur J-Y., Tabuteau G. (1996). Semantic Disambiguation in an Information Retrieval System. In Proceedings of Natural Language Processing and Industrial Applications NLP+IA/TAL+AI 96 Conference (pp.148-154). Université de Moncton, Canada.

Maybury, M. T., (Ed.) (1997). Intelligent Multimedia Information Retrieval. AAAI Press and MIT Press.

Miller, G. A., R. Beckwith, C. Fellbaum, D. Gross and K. J. (1990). Introduction to WordNet: an on-line lexical database. International Journal of Lexicography 3 (4), 235–244).

Monarch I., Carbonell J. G. (1987). CoalSORT: A knowledge-based interface. IEEE EXPERT, 39-53.

Paice C. D. (1991). A Thesaural Model of Information Retrieval. Information Processing & Management, 27 (5), 433-447.

Piatetsky-Shapiro G., Frawley W. J. (1991). Knowledge Discovery in Databases. AAAI/MIT Press.

Pollitt S. (1987). Cansearch: An expert system approach to document retrieval. Information Processing and Management, 23 (2), 119-138.

Qiu Y., Frei H. P. (1993). Concept Based Query Expansion. In Proceeding of Annual International ACM SIGIR Conference on Research and Developement in Information Retrieval (pp. 160-168). Pittsburgh PA, USA.

Smith P. J., Shute S. J., Galdes D., Chignell M. H. (1989). Knowledge-based search tactics for an intelligent intermediary system. ACM Transactions on Information Systems, 7, 246-270.

Sperberg-MaQueen, C. M., Burnard L. (Eds.). (1994). Guidelines for Electronic Text Encoding and Interchange. Chicago and Oxford: Tex Encoding Initiative.

Steinhart, E. (1995). NETMET: A program for Generating and Interpreting Metaphors. Computer and the Humanities, 28, 383-392.

Sutherland, K. (1990). A Guide Through the Labirint: Dickens's Little Dorrit as Hypertext. Literary and Linguistic Computing, Vol. 5, (4), 305-309.

Turtle H. R., Croft W. B. (1991). Evaluation of an inference network-based retrieval model. ACM Transactions on information systems, Vol. 9, (3), 187-222.

Vickery A., Brooks H. M. (1987). PLEXUS – the expert system for referral. Information Processing and Management, 23 (2), 99-117.

Voorhees E. M. (1994). Query Expansion using Lexical-Semantic Relations. In W. B Croft., C. Van Rijsbergen (Eds.), Proceedings of 17th Annual International ACM SIGIR Conference on Reseach and Development in IR (pp. 62-69). Dublin, Ireland.

Voorhees E. M. (1993). Using WordNet to Disambiguate Word Senses for Text Retrieval. In Proceedings of Annual International ACM SIGIR Conference on Research and Developement in Information Retrieval (pp. 171-180). Pittsburgh, PA, USA.

Woods W. A. (1997). Conceptual Indexing; A Better Way to Organize Knowledge.Sun Microsystems.