NAVILEX: Search and Navigation in a Semi-automatic Content Acquisition Legal Hypertext

ETTORE PIETROSAINTI, PAOLO MUSSETTO, GIANNI MARCHIGNOLI

1. INTRODUCTION

The goal of most legal work – seen as a process of text handling – is actually to combine different types of texts in an effective way. Statutes, regulations, cases, precedents, legal literature, contracts are examples of documents that may have to be investigated together in order to solve a legal problem or even to be able to understand the practical meaning of a legal rule. This well-known phenomenon can be described as legal rule fragmentation: the necessary information is often scattered in different documents or even in different data banks, and the links among the required pieces of information are difficult to establish. This problem, though particularly hard in a strongly text-centered field like the legal domain, is not limited to the legal area.

As a consequence, in the recent years a growing interest has been dedicated to the integration of search functions (typical of the Information Retrieval field) with navigation functions (typical of Hypertext systems). An effective combination of the benefits provided by the two models is expected to provide the best support tools for the localization of scattered information that is of interest for the user.

Among the different models that have been proposed in various fields, in the legal domain the interest for legal databases led to the adaptation for legal data [Agosti et al. 91] [Di Giorgi Nannucci 92] of a general two-level model [Agosti et al. 91a] providing a conceptual layer intended to improve the system transparency. Although a sound conceptual layer is a vital component of any effective model, we believe that a functional limitation of many existing models is the lack of context information, suitable to be combined with concepts to improve the retrieval accuracy (especially the precision component).

Besides this functional limitation, an economical problem which hampers the development of powerful models is the cost for the manual indexing/
authoring task. In the legal field the importance of the problem can be easily understood by considering the figures that express the rapid growth of the overall document collection: more than one million new cases and statutes per year, according to [Hafner 90].

The state-of-the-art Natural Language Techniques do not allow to envisage effective solutions to this problem on general domains. Nonetheless, in limited sub-languages the feasibility of an automatic mapping from texts to suitable information structures has been demonstrated in various fields [Liddy et al. 91] [Rama Srinivasan 93]. In the legal domain the Nomos project [Graziadio et al. 92; Giannetti et al. 92; Pietrosanti et al. 93] has demonstrated the potential of ‘shallow’ NPL techniques, as opposed to the traditional ‘deep’ techniques, for effective large-scale text processing in the legal domain.

Moving from the previous considerations, our work in the legal field has conformed to the following guidelines:

- Emphasis on the modelling of context characterization of concepts, to improve precision;
- Use of semi-automatic tools for the acquisition of the legal information components;
- Definition of a friendly interactive environment, referring to a book metaphor.

The resulting hypertext model is based on the close integration of the search and navigation dimensions, in a framework which offers multiple views (structural, conceptual and functional) over legal documents. The prototype system NAVILEX has been implemented, in order to experiment the potential for advanced applications of the model, dealing with a significant excerpt of the Italian Banking Regulations.

The main focus of this paper is on the description (including an overview of the relevant linguistic tools derived from the Nomos project) of the functionalities of the prototype, along with the underlying retrieval model based on a ‘legal electronic book’ metaphor.

The structure of the paper is as follows. Section 2 is devoted to the reference model for the integration of search and navigation functions on legal domains. The structural, conceptual and contextual dimensions are presented, together with the fundamental search and navigation functions made available by the model. The presentation provides also (2.4) a detailed overview and discussion of related approaches and an outline (2.5) of the basic assumptions underlying our approach to ‘shallow’ Natural Language Processing, including a brief description of the NPL modules.
The following section 3 describes the motivations and the functionalities of the NAViLEX prototype, including also (3.2) an extensive demonstration of the functionalities by means of an example of a user session.

2. **Fundamental Components of Legal Documents: A Reference Model for the Structural, Conceptual and Functional Dimensions**

The proposed reference model for the content representation of legal texts, is described in this section in terms of the multiple views (structural, conceptual and functional) provided over the legal texts. The presentation emphasizes the issues of the transparency of the model (relying on a ‘natural’ structure for the typical user who is familiar with legal text-books) and of the incrementality of the search process.

The *structural dimension* of the content representation includes the hierarchical organization of legal documents and the network established by the legal cross-references, which can both be used for navigation purposes allowing respectively direct access to a given text (through a table of contents) and a hypertext access through the ingoing or outgoing citations.

The *conceptual dimension* is based on the definition of complex linguistic terms (namely noun phrases) which constitute more effective content descriptors (as those typically included in the index of a book) with respect to word-based indexing [Croft et al. 91] [Evans et al. 91].

Another crucial dimension of the content representation is the *functional component*, which allows to associate to the concepts specific *functional roles* which constitute meaningful contexts for the user (e.g. the definiendum of a definition, or the subject of an obligation).

2.1. *The structural dimension*

Legal databases exhibit a general structure characterized by the well-known subdivisions of legal texts (e.g. sections, articles and paragraphs in statutes; chapters, sections and sub-sections in regulations). In addition to constituting the basic text cohesion device, this hierarchical organization is an essential aspect of the structural dimension of legal documents, as it provides also an ‘address space’ that is extensively used to make explicit reference, in a given legal document, to related parts of the same or other documents. The links established by these references make a complex cross-referring web of documents, that constitute the second component of the structural dimension.
Anticipating the description of the document base of the prototype (see the following section 3), Fig. 1 represents both components of the structural dimension.

The document base is composed of documents belonging to two related legal texts, whose hierarchical organizations correspond in the figure to vertical tree-like structures. The arrows illustrate the cross-reference links arising from two excerpts of the left-side text. The legal sub-domain is that of the Banking Legislation, which includes the Decree («Decreto») represented in the right-side structure. The articulation is typical of Statutes (issued by the Parliament): the Decree is composed of numbered Articles («Articoli») organized in Titles («Titoli») that are in turn made of Sections («Sezioni»), both indexed by roman numbers.

**Fig. 1. Structural Dimension of the Document Base (Regulation and Statute)**

In order to specify and give the correct interpretation of the whole Banking Legislation – that is scattered in several Laws and Decrees – the national Bank of Italy (Banca d’Italia) issues the *Regulations on Bank Surveillance («Istruzioni di Vigilanza»)*. These Regulations are represented by the structure on the left hand side, that is articulated in Chapters («Capitoli»), Sections and nested Paragraphs (for instance the upper text excerpt,
2.2. Providing a functional context for the conceptual dimension: modelling the norm-kernel and definitions as functional schemata

The investigation of legal concepts and their relations with other concepts is a crucial goal of much effort in legal work, aimed at identifying legally relevant items of knowledge and relevant relations between such items. The upper conceptual layer of the model (described in Fig. 2) represents the universe of possible usable terms and their relationships. For the purpose of our model, a concept represents a meaningful entity for the domain. Each concept is linked to the documents, in which concept instances are denoted by concepts anchors, that represent the linguistic manifestations of the concepts. In general, a set of possible linguistic expressions (i.e. the Concept Anchors present in the documents) denote a concept identified by a normalized linguistic term that represents the concept name (i.e. a noun phrase). For example, in the figure $C_a-A_1$ (e.g. «sales of goods») and $C_a-A_2$ (e.g. «to sell goods») are two anchors that identify two instances of the same concept $C_a$ (e.g. «sale of goods»).

Relations are established among concepts that are semantically linked. In the figure a generic hierarchical structure is depicted, that can be imagined as either a classification scheme or a complex thesaurus. Although we believe that the conceptual network is an important component for the complete model, in this paper the main focus is on the combination of concepts with relevant contextual information, aimed at improving the retrieval accuracy (especially the precision component).

The rest of the section is therefore devoted to the discussion of the contextual dimension, that is represented in Fig. 2 by particular frame structures. Since these structures express the particular function of the concepts in the context of the message communicated by the text, they have been named functional schemata.

The use of norm frames as a plausible method for the conceptual representation of legal knowledge has received large consensus in the legal theorist community. In a recent paper [van Kralingen et al. 93] this model is...
discussed in the context of Legal Knowledge Based system, in view of the particular task of the support to drafting of legislation. Although their goal – i.e. the complete representation of a norm – is far beyond the scope of our work, the representation schema that we propose shares with their model the norm-kernel (a long standing concept, cf. [Von Wright 63]). The norm-kernel is supposed to contain the essential information conveyed by a norm, answering questions like: who ought to do something?, what should he do?, etc. This leads to the consideration of the legal modality, subject, object and conditions of applications of a norm.

The legal modality determines the function of a norm, that is either an obligation (a command or a prohibition: ought or ought not respectively) or a permission (may). The subject of a norm is the person or institution to whom the norm is addressed. The conditions of application establish the circumstances under which a norm is applicable.

All of these components of the norm-kernel have been considered in the functional schemata of our model, because they play an important role as context qualifiers for the legal concepts. In the mentioned complete models for legal knowledge representation, the object of a norm is a fine-grained description of the act which falls under the scope of the norm, taking into account details such as the modality of the action, the setting of the action (e.g. spatial, temporal and circumstantial aspects) and the rationale of the action (causality, finality, intentionality). We take a different approach, and
reduce the complex activity components to a couple of slots: the action — i.e. the activity performed by the subject of a norm — and the object, that can be either a direct or indirect object of the action. In addition, we define a generic context role to take into account both the conditions of applications and the possible aspects of the action (e.g. temporal and spatial aspects).

We have considered so far the essential component of the content of legal texts, namely the propositions (norms) that express general rules, standard of behaviour and principles.

Another fundamental component of legal texts is represented by the definitions of legal concepts. The knowledge of the definition(s) of legal concepts is an essential pre-requisite for the correct interpretation of norms. In our model the definitions are represented by simple schemata that are intended to capture the essential functional roles played by the concepts that appear in the definitions. Using traditional naming conventions, we have thus identified the following functional roles: definiendum (the role played by the concepts that are defined in the definition); definiens (the role of the concepts that appear in the definition body); definition context (analogous to the corresponding context role for norms, generally includes concepts related to conditions of applications).

Following the previous considerations, our model takes into consideration the following fundamental functional schemata:

**SCHEMA TYPE: OBLIGATION/PERMISSION**

- **Subject**: <subject of the obligation/permission_
- **Action**: <activity performed by the subject of a norm_
- **Object**: <object of the action_
- **Context**: <relevant concepts for the context of applicability_

**SCHEMA TYPE: DEFINITION**

- **Definiendum**: <concept(s) to be defined_
- **Definien**: <defining concepts_
- **Definition context**: <relevant concepts for the conditions of applicability_

The choice of simple one-level functional schemata has a twofold motivation. First, our goal is to provide a simple conceptual and functional view
of the document content, to be directly used for intelligent retrieval purposes. Complex schemata with a large number of nested slots would be difficult to manage for the user. The second motivation is related to the strategic goal of automatic acquisition of the content representation. While we have experimented (as shown in section 2.5) the feasibility of shallow Natural Language Processing techniques capable of extracting the simplified schemata directly from texts, the automatic acquisition of a fine-grained representation is hampered by the state-of-the-art limitations of full-fledged NLP.

To clarify the previous considerations, we present a brief excerpt extracted from Chapter L, Section I, Paragraph 2.c of the mentioned Regulation:

«Should the reorganization process give origin to a group of credit institutions as mentioned in Title VII of the decree ..., the project, ..., ought to describe:
– the organizational structure ...;
– the articles of association of the holding company as mentioned in art. 25 of the decree (see Section III)...»

The relevant concepts are highlighted, while the underscored words emphasize the recurrent cue phrases that are used for the acquisition of functional schemata (see section 2.5 below). The corresponding representation in terms of concepts assigned to the relevant roles of a functional schema is the following:

SCHEMA TYPE: OBLIGATION

Subject _
   project
Action _
   describe
Object _
   organizational structure
   article of association
   head holding company
Context _
   reorganization process
   group of credit institutions
   Title VII of the decree

2.3. Basic search and navigation functions over the reference model

This section is aimed at outlining the fundamental functions provided by the model for search and navigation support, concentrating on the functions related to the conceptual and contextual dimensions.
We just mention the typical functions related to the structural dimension: they include the capability to navigate over the network of cross-references and to traverse the hierarchical structure. These functions are described together with the prototype in section 3.

The functions related to the conceptual and contextual dimensions make reference to the following basic entities of the model previously outlined (see Fig. 2):

- Concepts
- Concept anchors (instances)
- Functional schemata
- Functional roles

**Concept based document selection**

Given a concept \( C \), the output of the function is the set of documents which include instances of the given concept. In terms of the model, the output document set is calculated by traversing all the links from the given concept to the associated concept anchors in the documents.

For example, the concept \( C_a \) selects both the documents represented in Fig. 2. This functionality has been modeled in the NaviLex prototype as a search over the index of the book.

**Context qualification of the instances of a concept \( C \) in a document \( D \)**

Given a concept \( C \) whose instances are present in a document \( D \), the function returns the available information about the functional roles associated to the concept instances.

For example, given the concept \( C_b \) and the Document \( D_b \), the function returns the role Role2 of the right hand schema. This function can be used in the prototype to know the role of a given concept in a document selected via the index.

**Context based concept browsing**

Given: 1) a document set \( DS \) (possibly the entire document collection); 2) a functional role \( FR \),
the function returns a set of concepts whose instances play a role \( FR \) in some document belonging to the document set \( DS \).

For example, given the two-document collection of Fig. 2 and the functional role Role1, the function returns the concept set \( \{C_a, C_b\} \). In case of a definition schema (Role1 = Definiendum) \( \{C_a, C_b\} \) would be the set of concepts for which a definition is present in the given document set.
• **Document selection based on a concept’s functional role**

  **Given:**
  1. a document set DS (possibly the entire document collection);
  2. a Concept C;
  3. a functional role FR;

  the function returns the document subset of DS that include instances of the concept C playing the functional role FR.

  For example, given the two-document collection of Fig. 2, the concept C, and the functional role Role1, the function returns the document set {D2}.

2.4. Other hypertext retrieval models

  The purpose of this section is to discuss the distinctive features of our approach in comparison to other hypertext models and systems, that have in part been mentioned in previous sections. As observed in [Arents and Bogaerts 93], all the models recently presented have in common the separation between the *document space* (the documents in the hypermedia system) and the *index space* (the indices that characterize these documents).

  Our model conforms to this general viewpoint and in particular can be closely related to the *explicit* model presented in [Agosti et al. 91] and referred to in [Di Giorgi Nannucci 92], where the document and index space are respectively indicated as *hyperdocument* and *hyperconcept*. The hyperdocument, defined as a network of structural links combined with the network of reference links, corresponds closely to the structural dimension of our model. The conceptual dimension of our model matches the purpose of the hyperconcept, whose task is to handle the semantic structure of concepts used to describe the contents of document collection. The use of a rich semantic structure (including *indexing links* connecting thesaurus nodes to the documents as well as *classification links* used to aggregate documents according to classification criteria) is proposed also in [Aboud et al. 93].

  With respect to these models, a crucial additional feature of our model is the consideration of the context information, provided by the functional schemata (contextual dimension).

  The contextual information is taken into account in [Arents Bogaerts 93] in connection to what they call *semantic hyperindices*, that rely heavily in the use of thesauri to support browsing search. The example «Definition of pitting-corrosion of titanium» shows how they use the context «definition» to qualify the occurrence of the concept «pitting-corrosion of titanium». In their work much emphasis is on a sophisticated description of concepts, which are embedded in a network that can be traversed both
vertically and horizontally with a link navigation mechanism called ‘broad-
button’.

In [Nanard et al. 93] a model is described that explicitly considers the
qualification of concepts by means of contexts (examples of contexts are
«definition», «general rule», etc.), thus allowing contextual access to technical
documents. They also mention the issue of the semi automatic elicitation
of information from the initial document base; improvement about this
point is said to be under way.

With respect to the mentioned models that take into account the context
information, a distinctive feature of our model is that the functional sche-
mata are actually structured contexts, in which the user can specify not
only a context type (for example the «definition» schema type) but also the
functional roles associated to the concepts of interest.

Our notion of structured contexts is also related to the idea of segmented
database that is discussed in [Rau Jacobs 91], where the differentiation of
keywords into segments allows to distinguish – in a constrained domain of
commercial news – companies mentioned in passing from those actively
involved in mergers or other events, and locations of companies from loca-
tion of stories. This capability, that is achieved using sophisticated Natural
Language Technology, is analogous to the possibility to distinguish (see the
following section 3.2) between the occurrence of a concept as the definiendum
of a definition and the occurrence as the subject of an obligation.

In view of our research objective related to the extraction and use of
context information, important insights and results are given by [Rama and
Srinivasan 93] who have reported on an investigation carried out on medical
abstracts, in order to show how the qualification of keywords with their
conceptual roles in a text can be used to derive a meaningful text-repre-
sentation scheme. In addition to the text-grammar approach based on observa-
ble regularities in the structure of documents, that is related to our techniques
for the recognition of functional schemata (see section 2.5 below),
their paper presents interesting experimental results. They have studied the
role distribution of keywords, and found out that keywords exhibit role
variation across abstracts, making the claim that this variation can be
potentially exploited to make retrieval more precise. This result supports
our experience with the use of functional schemata in the NaviLex proto-
type. As far as the use of the text-grammar roles is concerned, their approach
is in contrast to ours, as they believe that their typical frame (that is actually
much more complex than our functional schemata) would be too much diffi-
cult for users to be adopted as a query device. They propose the alternative
strategy of using the text grammar as a selection mechanism for extracting
the most important information for the abstract. This information is then used for retrieval strategies based on the notion of spreading activation.

2.5. Linguistic tools for the semi-automatic acquisition of the reference model components

Legal texts are written according to fixed conventions, some of them explicitly stated and recommended, others just de facto standards: typographical layout, formal and recurrent expressions and a specialized vocabulary make legal texts an ideal test-bed for experiments in the «sub-language» area [Liddy et al. 91] [Rana Srinivasan 93].

The issue of automatic knowledge acquisition from legal texts, relying on their highly structured nature, has been a major goal of the Esprit II Project n.5330 «NOMOS: Knowledge Acquisition for Normative Reasoning Systems» [Graziadio et al. 92; Giannetti et al. 92; Pietrosanti et al. 93]. The project was actually divided into two parallel sub-projects, dealing respectively with legal texts in French and Italian: the approach described in this paper deals with the ILAM (Italian Language Acquisition Modules) processors, that have been applied to the Italian Value Added Taxation Law.

In Nomos a specific set of linguistic tools is devoted to the extraction of a ‘shallow’ content representation layer, including several structural, conceptual and functional components of legal texts that, although not adequate for legal decision support applications, can be effectively used for purposes of ‘intelligent’ navigation and retrieval on legal texts. A crucial point is that in this case the relevant automatic acquisition modules rely on ‘partial’ NLP techniques that do not require the most powerful, computational aspects of natural language processing but are based on recurrent legal text peculiarities that make them suitable to be applied to large Text Databases.

Generally speaking, our model can be situated at the knowledge representation level which has been named in [Nanard et al. 93] macroscopic semantics, being far more detailed than simple indexing by weighted keywords, but far less detailed than a conceptual-graph based description needed for complete text understanding.

The Cross-reference Specialist is a linguistic processor aimed at detecting and formalizing the linguistic expressions that denote legal references. This formalization can be used to populate the database that represents the web of cross-references.

The references are usually indicated by means of complex but regular linguistic expressions dealing with the links to other parts of the law. The input of the processor is a normalized encoding of the Legal Text; the
output is an enriched encoding of the Text, in which any internal or external reference is represented in the cross-reference record format, that can result in the tagging needed to generate an hot-key useful for navigation.

The main knowledge source used by the tool is the Law-Reference Grammar, that represent the knowledge of the possible surface structures of the cross reference linguistic expressions. In the grammar linguistic phenomena such as conjunctions and anaphoric references are also handled.

The semi-automatic acquisition of concepts is supported by a tool – the Concept Processor – aimed at performing a morpho-syntactic normalization of the surface linguistic expressions, obtained through the syntactic recognition of the proper conceptual units (i.e. noun phrases).

The objective of the analysis is to detect, using a «shallow grammar», the surface syntactic links between word pairs or triples (rather than a ‘deep’ parse tree of the sentence).

For example, in the sentence «The transfers of goods are considered done in the state territory if they concern real property» the following noun phrases can be recognised:

- N_prep_N(transfers, of, goods)
- N_N(state, territory)
- N_Adj(real, property)

This surface parsing allows reduced computational complexity and flexible adaptation of the grammar to different linguistic domains. The grammar currently includes more than 20 rules, including «skip» rules. Skip rules are rules which allow discontinuous constituents of a Noun Phrase to be linked. In addition to the skipping and distribution over conjunctions, a nominalization mechanism has also been implemented, allowing to transform suitable Verb Phrases into the corresponding Noun Phrase (e.g. «to sell goods» becomes «sale of goods»).

The Functional Processor allows to classify the input sentence as an instance of a specific Functional Schema (e.g. a Definition Schema), whose structure is obtained by decomposing the input into text-segments that are given the appropriate functional roles (e.g. the definiens or the definiendum in a definition schema).

The processor relies upon the expectations established by the recognition of recurrent text patterns (whose number is manageable due to the uniformity of legal texts) exploiting a set of ‘semantic markup symbols’ (cue phrases, e.g. definitional verbs) that constitute the ‘linguistic backbone’ of the functional schemata (a similar approach has been adopted in [Moulin Rousseau 90]).
In addition to the basic segmentation due to the existence of markup symbols, additional mechanisms have been devised in order to cope with the lack of suitable cue phrases. The first strategy relies on heuristics based on syntactic information: in some definition patterns with contiguous segments, the splitting point is identified by interpreting the whole string as a NP + NP sequence, as in the following example, dealing with the definition of state territory:

«is considered» <definition verb>  
«state territory» <definiendum slot> NP  
«that subject to ...» <definiens> NP

In most cases, for the recognition of the sequence it is sufficient to search for some closed-class splitting words such as determiners and pronouns; in the general case, the concept processor can be called to identify the couple of contiguous Noun Phrases.

The second strategy is based on a ‘learning’ mechanism relying upon the legal sub-language lexical uniformity (i.e. the typical reiterativity of the juridical style); in this case the lack of suitable markup terms can be overcome by resorting to some previous occurrence of the same slot-filler combination in the current context (e.g. a Law Article).

3. THE NAVILEX PROTOTYPE: A BOOK METAPHOR FOR NAVIGATING IN A LEGAL HYPERTEXT

In order to validate the adequacy of the model, we designed and developed a prototype addressed to a specific user, in the context of the banking legislation. The target users are expert legal drafters working in the legislative department of the Bank of Italy (Banca d’Italia), in charge of the drafting and maintenance (with respect to the evolution of the relevant statutes) of the Regulations.

The preliminary task analysis indicated the requirement for an intelligent retrieval aid, possibly capable of supporting the user in the detection of possible inconsistencies. As the regulations constitute a fairly dynamic document set, the legal fragmentation is a real problem for the draftsman who is in charge to modify the Regulation by preserving its consistency.

In addition to the mentioned functional requirement, we considered that the target user was not familiar with the use of computerized data banks, whilst was obviously acquainted with the consultation of traditional paper books.
These considerations suggested to organize the information components defined in the reference model in a friendly environment, based on a ‘legal electronic book’ metaphor.

The issue of converting a textbook to hypertext has also been addressed in [Rada 92], in which the author reports that users value book metaphors. A distinction is also made there between first-order links and second order links. The former are those connecting outline headings, bibliographic citations, indices, footnotes and cross-references, while the latter are links that are not explicit in the text. In [Rada 92] the focus is on experimenting the conversion of a textbook into four different hypertext systems, making the assumption that the document is already available in electronic form with suitable logical markup representing the first-order links. In our case, the only available first-order links were the text pages (sections) separated by section headings, while all of the remaining first order (i.e. structural) and second order (i.e. conceptual and functional) links were not available.

The main focus of our experimentation was on the usability of the information components and on the interaction paradigm. Nonetheless it is important to emphasize that each information component used in the prototype can be extracted semi-automatically (i.e. a certain amount of manual intervention and validation is currently necessary) with the techniques first experimented in the processors of the NOMOS Esprit project described in section 2.5.

Following an explicit user indication, the prototype’s document base has been centered on two particularly complex Chapters of the Regulations, namely the Chapter 50 (Reorganization processes) and Chapter 52 (Groups of credit institutions), whose size is about 20 paper pages. In addition, as indicated in section 2.1, the system allows cross-reference navigation from the regulation to relevant articles of the Decree.

3.1. Outline of the prototype functions

A major goal for the prototype was to realize a ‘natural’ environment for the user who is acquainted with book reading. Therefore, adopting a book metaphor, the system includes:

a) Searchable Table of Contents (relying on the hierarchical structural component);

b) Two-level conceptual index (relying on the conceptual organization);

c) Navigation over the reference links (based on the cross-referring web of documents).
Building over this ‘book-oriented’ use of the information components, the system offers the three functionalities described in the previous section 2.3, based on the functional schemata context representation:

d) Context qualification of the instances of a given concept in a document;

e) Context based concept browsing;

f) Document selection based on a concept’s functional role.

The NAVILEX prototype has been developed on an IBM PC workstation, using Toolbook in Windows 3.1 environment. The structured data have been represented with a dBaseIII database, accessed via an available DLL supplied with Toolbook.

The next section is devoted to a description of both the environment and the functionalities, by means of a comprehensive example simulating a user session.

3.2. Example of user session

We assume that our user (U) is interested in the concept of «capogruppo» (holding company), with particular reference to the exact definition of the concept and to the related obligations.

Fig. 3
He begins by accessing the table of contents (TOC, in the interface "Indice Generale"). The interface displays (see Fig. 3) a window with the list of available sections of the regulations, properly indented to render the hierarchical relations. Whenever the title length should exceed the field dimension, U can double-click an item in order to get a pop-up window with the full title.

In the lower part of the window, the user can type the concept of interest (i.e. ‘capogruppo’) to retrieve all the paragraph descriptors whose title includes the concept. Many paragraphs have fairly long titles, that can be considered as (very) short abstracts, so this search can be expected to select (with low recall) a subset of the relevant documents.

### 3.2.1. Cross-reference navigation environment

In order to describe the navigation environment, we assume for the moment that U is interested in reading the document whose titles ‘capogruppo’ (Cap. LII Sez. II Par. 1.), so by clicking the relevant item of the TOC he gets to the window shown in Fig. 4.

This window displays (in a scrollable text field) the document text, that contains a definition of the concept ‘capogruppo’. The interface offers a number of functions to navigate the document base. The first is the sequen-
tial scanning of the regulation database: the relevant icon (on the left hand lower corner) evokes the activity of turning over the pages of a book, and the two triangular buttons allow to scan the book forward or backward. The next icon (depicting a set of playing cards to suggest a selection mechanism) corresponds to a constrained sequential search, limited to the document set selected with the previous search in the TOC.

To navigate the cross-reference network, the interface provides ‘hot keys’ in correspondence to the position of the cross-references in the text (outgoing links). For example, if U positions the mouse cursor over the string «cfr. par. 1.1» («see paragraph 1.1») the cursor changes its shape, denoting the presence of an ‘hot key’: if U clicks the mouse, the system displays the text corresponding to paragraph 1.1.

The dual of this operation is the navigation through the ingoing links, namely the navigation towards the documents that contain reference to the current document. By clicking to the ‘target’ icon, U can get the list of the document descriptors that are the source of the ingoing links.

The button ‘Back’ allows to return back along the path traversed during the navigation; the remaining two icons allow respectively to access directly a given page (without passing through the TOC) and to store in the network user-defined links or bookmarks.

3.2.2. Functional context-aided conceptual search

After the previous digression concerning the navigation environment, we come back to the user’s information need concerning ‘capogruppo’, at the point where he had obtained (in Fig. 3) a list of five documents. We assume U cannot accept the low recall expected from the search over the TOC, so he gets to the Index («Indice Analitico») shown in Fig. 5. The index is intentionally structured as a conventional two-level index of a paper book. The user selects the entry of interest (i.e. ‘capogruppo’) and the system retrieves the list of documents displayed under the item ‘Pagine Recuperate’ (Retrieved Pages). The list actually contains 13 items, suggesting a low precision rate (the documents associated to the index concepts are not selected by means of particular relevance criteria). In the case of the prototype’s small document base, the user could browse the documents of the list, but in a real system the dimension of the list would discourage this operation.

The availability of the context qualifications function can help, providing a sort of ‘functional preview’ of a document. In the figure the user has selected the document Cap. LII. Sez. 1, and pressed the button ‘Schemi’ (Schemata). The two lists on the right hand side inform the user that in the
**FIG. 5**

![Diagram](image)

**FIG. 6**

![Diagram](image)
selected document the concept ‘capogruppo’ acts both as the definiendum of a definition and as the object of an obligation.

The User does not want to repeat this operation for all the documents, so decides to select directly the documents relying to the interesting functional roles for the required concept.

The status of the window displayed in Fig. 6 has been reached after the following sequence of User operations. The user was interested in locating the definitions of ‘capogruppo’, so has selected the item ‘Definizione’. The system has thus displayed the frame-like structure constituted by the three functional slots of a definition schema, namely Definiendum, Definiens and Contesto (Context). By clicking on the ‘key’ button related to the Definiendum slot, the user has obtained the list (‘Concetti Selezionabili’) of the concepts belonging to the Definiendum role in some definitional schema known to the system (this operation corresponds to the function previously named Context based concept browsing). The user has selected from the list the concept ‘capogruppo’, and the system has retrieved three documents (including the LII-II-1 displayed in Fig. 4). In this way the user has been able to identify directly the three documents of interest (out of the 13 retrieved with the index) improving considerably the precision of the search.

From the same window the user can ask the system to consider the

---

**Fig. 7**

- **Accesso Diretto per Schema Funzionale**
- **Definizione**
- **Obbligo**
- **Possibilità**
- **Soggetto**: capogruppo
- **Azione**: key
- **Oggetto**: key
- **Contesto**: key

**Concetti Selezionabili:**
- amministratori enti controllati
- amministratori società controllate
- capogruppo
- ente creditizio pubblico
- progetto ristrutturazione
- società bancaria capogruppo
- società controllate

- **Pagine Recuperate:**
  - Cap.LII.Sez.III: 2
  - Cap.LII.Sez.III: 1
  - Cap.L.III.Par.6.: 1
  - Cap.L.II.Sez.IV.Par.4.1: 1
  - Cap.L.II.Sez.IV.Par.4.2.4: 1
Obligations concerning ‘capogruppo’. The situation is illustrated in Fig. 7, where the right hand list displays the descriptors of the 5 documents that contain obligation schemata where ‘capogruppo’ is the subject of the obligation. The subsequent Fig. 8 shows the possibility of incremental retrieval offered by the functional schema search paradigm that, as mentioned before, constitutes a distinctive feature with respect to other models. The user is supposed to have requested – by pressing the ‘key’ button associated to the role ‘Azione’ (Action) – the list of the action concepts which are present in the schemata in which the subject of the obligation is ‘capogruppo’. By selecting the concept ‘comunicare’ (to communicate) he has narrowed the search result to the list of three documents displayed in Fig. 8.

4. Conclusions and Future Research

Starting from a critical evaluation of the benefits and limitations (under both a functional and economical viewpoint) of the existing models which integrate search with navigation functions, we have presented a reference model which takes into account context information in order to improve the retrieval accuracy. The model has been experimented in the NAVILEX
prototype, – dealing with a significant excerpt of the Italian Banking Regulations – that has been presented including also an overview of the linguistic tools devoted to the semi-automatic extraction from texts of the information structures considered in the model and used in the prototype.

A detailed evaluation of the prototype is ongoing. The preliminary comments from the target users (i.e. drafters of the Italian Banking Regulations) are quite encouraging. The user is satisfied with the interaction model, and recognizes the potential of the system as a support tool in the task of detecting possible inconsistencies in the legal system.

Future developments are envisaged for all the aspects dealt with in the paper, namely the model, the hypertext system and the acquisition tools.

We are investigating an extension of the model, aimed at the conceptual qualification of the cross-reference links arising from second level legal texts (i.e. cases and regulations) towards the statutes. The goal is to improve the navigation from a given statute to the related (e.g.) cases, relying to concepts that properly describe the reason of the link, thus providing an indication about the possible interest of the link for the user problem.

Concerning the evolution of NaviLex into a system, we are planning to enhance the conceptual dimension, by using a fully-fledged thesaurus structure. This should improve the overall recall of the search functions, thus matching the parallel improvement provided on precision by the contextual information. Concerning the interface, an important aspect would be the enhancement of the navigation of the hierarchical structure, with the ability to fold/unfold the outlines.

We are also making experiments with commercial IR systems capable of the definition of structured contexts (e.g. using SGML tagging) in order to verify the possibility to integrate context-based searching into standard IR environments.

The last remark is about the enhancement of the linguistic indexing tools for the semi-automatic extraction of structured information. The work started with the NOMOS project is being continued by a research activity funded by the FINSIEL Group; a large-scale evaluation of the tools is ongoing, in order to assess the adaptation effort due to sub-domain linguistic shift (changes in the lexicon, in the recurrent stylistic cues, etc.).

References


