Contextualizing the Information Space in Federated Digital Libraries

M.P. Papazoglou, J. Hoppenbrouwers
Tilburg University/Infolab
PO Box 90153, NL-5000 LE Tilburg
The Netherlands
{mikep,hoppie}@kub.nl

Abstract

Rapid growth in the volume of documents, their diversity, and terminological variations render federated digital libraries increasingly difficult to manage. Suitable abstraction mechanisms are required to construct meaningful and scalable document clusters, forming a cross-digital library information space for browsing and semantic searching. This paper addresses the above issues, proposes a distributed semantic framework that achieves a logical partitioning of the information space according to topic areas, and provides facilities to contextualize and landscape the available document sets in subject-specific categories.

1 Introduction

The most important problem specific to digital libraries with spatial distribution is the federation problem: making distributed collections of heterogeneous documents appear to be a single (virtually) integrated collection. Such a federated digital library (FDL) addresses a *narrow* and *specific domain area*, e.g., Biomedicine, Computer Science, or Economics. Usually the access to an FDL is through the World Wide Web on the basis of specialized search engines and browsers.

FDLs are faced with at least two major technical challenges. Firstly, document handling is hard as there is a large number of documents with differing type, structure and terminology. Secondly, due to the large number and variety of documents available, unless classification schemes are employed – so that document sources can be indexed in different ways and different levels of detail – distributed searching cannot be feasible [15].

The typical search question posed to a digital library is a specific search query, which matches some documents over the entire FDL. For example, a possible query for an FDL in Economics may be *"find all documents on reduction of the state deficit by tax increase"*. Most recent approaches to World Wide Web (WWW) querying [2, 8] concentrate only on *keyword retrieval*, viz. queries on *semantic content*. They naively assume that the user (or search engine) is explicitly aware of the structure, semantics and vocabulary differences of the Web data to be queried. However, due to the multiplicity, complexity, and terminology fluctuation of the data available, unless users are explicitly aware of both the structure and semantic nature of the data available, such querying cannot be successful.

Practical studies have shown that there is a critical mismatch between a user's and the Web's vocabulary [15]. Picking the right terms depends on how intimate searchers are with the vocabulary use in documents they wish to retrieve. The way that the user interacts with an FDL, by means of a browsing mechanism and a specialized search engine, may be described as follows. First the user seeks to understand her information needs by trying different terms and alternatives; once the user has found a potentially matching/interesting term then she may decide to learn more about the semantic context of this term. Finally, once the user has satisfied herself that she has located the term she is seeking then she retrieve the document (semantically) matching this term.

Of particular interest to such types of queries are *subject gateways*. These are facilities that allow easier access to network-based information resources in a defined subject area [7]. Subject gateways offer a system consisting of a database and various indexes that can be searched through a Web-based interface. Each entry in the database contains information about a network-based resource, such as a Web page, Web site or document. Typical examples of subject gateways are: the Social Science Information Gateway (SOSIG) [16], which incorporates a thesaurus containing social science terminology, and the Organization of Medical Networked Information (OMNI) [11] which allows users to access medical and health-related information. The key difference between subject gateways and the popular Web search engines, e.g., Alta Vista, lies in the way that these perform indexing. Alta Vista indexes individual pages and not resources. For example, a large document consisting of many Web pages hyper-linked together via a table of contents would be indexed in a random fashion. In contrast this subject gateways, such as OMNI, index at the resource level, thus, describing a resource composed of many Web pages in a much more coherent fashion.

Our research focuses on methods that interactively provide the user with a conceptual partition of the information space in FDLs into meaningful subject areas where different kinds of term suggestions can be used to enhance retrieval effectiveness. We refer to this logical organization as the *Topi*c-based Document *C*lustering *A*rchitecture (TopiCA). In TopiCA, the structure of an FDL resembles that of a massive semantic network where semantic links are created between document resources that are topically related. In line with the spirit of subject gateways TopiCA indexes resources, viz. documents and related document collections. Indexes are searched by routing from topic to topic (versus document to document as is current practice with Web database approaches) until the appropriate documents are found. TopiCA serves a multiple function: it helps users avoid search terms not used by indexers, while suggesting closely associated terms which have distinct and precise meaning within a number of a related collection of documents. Thus, users are allowed to fully articulate their needs in terms of meaningful queries against the FDL information space. Querying that structure of document data can subsequently be accomplished using query languages for the Web in the database style [2]. Other research activities which have influenced our ideas are work on the InfoHarness project [5] and work on Bayesian inference logics for contextual classifications [12].

2 Related Work

In the following we will describe similarities and differences with some of the information retrieval (IR) techniques that have influenced our work.

In IR-oriented applications, an important distinction is that between the document space and and some form of index space [1, 14]. Terms in the index space are descriptors for media items in the document space, e.g., text passages, images, etc. Many IR systems have been based on statistical analysis of terms automatically extracted from free text [3]. With these systems the index space is automatically created and a vector space similarity coefficients measure degrees of (semantic) match between queries and media supplied items, or between two media items. An alternative scheme to the traditional IR approach can also be used which is based on manually created index spaces where semantic relationships between index terms exist. It is also possible to combine these two IR approaches with a thesaurus used to expand query terms [6]. Our approach is partly based on the latter approach which it expands by categorizing documents, by semantically partitioning the information space, and by providing category-specific navigation and querying capabilities. These ideas are based on traditional IR clustering techniques described below.

In most clustering IR techniques the strategy is to build a static clustering of the entire collection of documents and then match the query to the cluster centroids [18]. Often a hierarchical clustering is used and an incoming query is compared against each cluster in either a top-down or a bottom up manner. Some variations of this scheme were also suggested in which a document that had a high similarity score with respect to the query would first be retrieved and then would be used for comparison to the cluster centroids. However, if a query does not match any of the pre-defined categories then it would fail to match any of the existing clusters strongly. As a remedy to this problem previously encountered queries are grouped according to similarity and if a new incoming query is not similar to any of the cluster centroids it might be instead similar to one of the query groups, which in turn might be similar to a cluster centroid. Our clustering techniques, although employing many of the traditional IR clustering algorithms, follows a different approach. First documents are sorted and tied to their high-level centroids (called generic concepts in this paper) and then interactive tools are provided for the user to expand or narrow her/his context and disambiguated her/his terms (via navigation through a lexical network). Once the centroid that contain these terms is determined then queries can be issued against its underlying document sources. This results in semantically disambiguated queries which avoid the zero-hit problems in traditional IR techniques.

3 Logical Organization of Documents in Federated Digital Libraries

To add a document to a (digital) library, we must index important document terms for efficient retrieval of the document. This is standard practice for all libraries. Surrogates of the documents in a digital library – called *document index records* (DIRs), or *meta-data* – are created by professional catalogers and indexers. The concept of meta-data typically refers to information that provides a brief characterization of the individual information objects in a DL and is used principally in aiding searchers to access documents or materials of interest [17].

The general idea is to group aggregated document index records together (via their respective schemas) into a topically-coherent group, and present textual summaries and a common structured vocabulary of topical terms to searchers for interaction. Only in this way we can allow tools and searchers to selectively access individual document aggregations while ignoring others.

The goal behind TopiCA is to provide tools that organize documents according to meaningful categories within a broad topic , e.g., Economics, and help users negotiate the structure and semantics of their information items against the explicit and implicit characterizations that have been extracted from a vast document space. The information analysis tools that are under development combine lexical analysis and navigation techniques with ontology-based categorization. TopiCA provides user interfaces for visual display of subject partitionings, organizes the information sources accordingly and gives the users the means to identify topics of interest to them. It also provides the terminology based context against which users can map their own keywords in order to retrieve documents that may contain semantically related terms.

To exemplify the TopiCA environment we use a comprehensive example from an *Education & Training* FDL connecting educational and training service providers, publication providers, accreditation, and government agency document servers. This FDL, which is based on the TopiCA architecture, shown in Figure 1, is in a position to provide a conceptually holistic view and cross-correlate information from the multiple document servers. We will describe this process in two broad phases.

Firstly, we employ a database like schema, the DIR schema (or *meta-data schema*) to describe the structure of the DIRs and specify how distinct sets of index-records and their terms can be logically aggregated to describe a particular subtopic. For example, aggregation of meta-data schemas which abstract documents containing information about courses, committees, accreditation-processes and so on, may represent subtopic such as a Accreditation. These terms represent objects in the individual meta-data schemas which in their turn may contain attributes. For instance, attributes such as course-name, credit-points, duration, etc would be contained in a course object which is part of the Accreditation subtopic. This process corresponds to steps 1 and 2 in Figure 1. Semantically related subtopics such as for example Accreditation and Enrollment-Program are also connected into a higher-level construct, see step-3, which we call a *Generic Concept* GC, e.g., Education. GCs thus represent semantically related DIR clusters (via their respective meta-data schemas) and form topically-coherent groups that unfold descriptive textual summaries and an extended vocabulary of terms for their underlying documents. A GC is thus a form of a logical object (a kind of *a contextualized abstract view* over the content of large semantically related document collections) whose purpose is to cross-correlate, collate, and summarize the meta-data descriptions of semantically related network-accessible data.

Secondly, to resolve terminology mismatches and semantic drifts between disparate index terms, topical synoptic knowledge and a standard vocabulary for term suggestions is supported by each GC. A GC materializes a class hierarchy depicting all terms within the topic sampled by the GC, e.g., Education. Each GC is characterized by its name and the context of its terms (term hierarchy and term descriptions) for each specific topic. Terms within a GC are shown to have a distinct meaning (sense) and context. This concept space consists of abstract descriptions of terms in the domain, term senses, relationships between these terms, composition of terms, terminology descriptions, hypernym, hyponym, antonyms-of, part-of, member-of (and the inverses), pertains-to relations, contextual usage (narrative descriptions), a list of keywords, and other domain specific information, that apply to the entire collection of members of a GC. Moreover, it includes other useful details such as: geographical location of documents, access authorization and usage roles, explanations regarding corporate term usage and definitions, domains of applicability, charge costs, and so on. Hence, the GC structure is akin to an associative thesaurus and online lexicon (created automatically for each topic category). Thesaurus-assisted explanations created for each subject-based abstraction (GC-based information subspace) serve as a means of disambiguating term meanings, and addressing terminology and semantic problems. Therefore, the GC assists the user to find where a specific term that the user has requested lies in its conceptual space and allows users to pick other term descriptions semantically related to the requested term. The GC entries, at the moment, are partially generated by the lexicographic framework WordNet [9] that currently underlies TopiCA. After semantic disambiguation the users are provided with the set of documents that contain the selected terms.

In summary, the TopiCA structure supports semantic reconciliation of autonomous interconnected document sources as it helps the users understand what information is available through the network; helps them categorize and configure their information demands on the basis of the information available to them; and assists them to semantically disambiguate their specified terms against those provided by sets of documents in a federated digital library.

3.1 An Example

The *Education & Training* FDL comprises a set of GCs such as Education, Training, Literature & Publications,

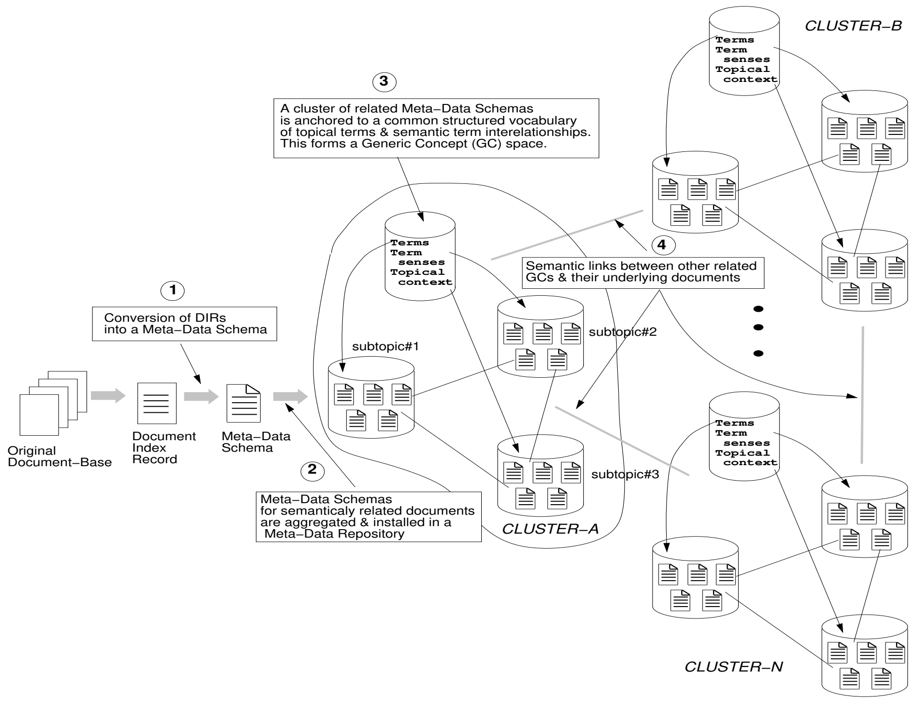

Figure 1: Connecting meta-data schemas and forming the Generic Concept (GC) space.

Employment, and so on. The topic-areas, described by the GCs, are interconnected by weighted links to make the searches more directed. When dealing with a specific subtopic such as Accreditation DIRs are not only able to source appropriate information from remote document-based on the same topic but also to provide *matching* information about enrollment programs, training schemes, research activities and publication data.

Overall a networked digital library system (representing a narrow domain) may be viewed in terms of three logical layers. The bottom layer corresponds to the document collections (document-base in Figure 1). The middle layer represents the meta-data schemas associated with the documents. The top most layer corresponds to the *concept space* (GC) layer. This three-tier architecture is the key ingredient to information elicitation in federated DLs. It generates a semantic hierarchy for document terms in layers of increasing semantic detail (i.e., from the name of a term contained in a document index, to its structural description in the meta-data layer, and finally to the concept space layer where the entire semantic context – as well as patterns of usage – of a term can be found). Searches always target the richest semantic level, viz. GC layer, and percolate to the schema layer in order to provide access to the contents of a document-base, see section-4.

Currently, in TopiCA human indexers assign "aboutness" to meta-data schemas (and consequently their underlying documents). Human indexers decide the degree of relatedness between meta-data schemas and GCs. Meta-data schemas belonging to a particular GCs, e.g., courses which belongs to the GC Education, normally link to it strongly. By *strongly linking* to a certain GC, e.g., with a weight 10/10, meta-data schemas associate with each other and thus inter–topic (and document) organization is achieved implicitly. Each of the meta-data schemas may also link (via its own GC) less strongly to other GCs which have their own associated cluster of DIR documents. The resulting GC structure forms a massive dynamic network, resembling a cluster–based *associative network* (a variant of semantic networks that uses numerically weighted similarity links). In this way, the entire set of relationships for a document collection within a specialized subtopic, e.g., Training or Employment, is organized into a lexical network in which the vocabulary items are represented as nodes and the semantic relationships within them are represented as links [10]. Moreover, each lexical network points to other related networks, e.g., Training links to Employment, depending on their degree of relatedness Figure 3(b).

4 Interacting with the GCs

An example of the GUI for some of the the terms included in the educational GC is given in Figure 2. Here, we assume that a user who searches the entries in the educational GC is interested in the term course and wishes to gain more insight into its semantic context in order to formulate a query at a later stage. The first step after entering the term is to choose the *senses* from the list the GC lexicographic substrate provides in the form of a menu (not shown in this figure due to reasons of brevity). The sense number returned is then associated with the term. In this particular example although the term course has eight senses (meanings) once the domain of discourse is limited to study (education) only one of the eight can occur. Figure 2 illustrates an expansion of the specific term chosen. This figure shows how the GC provides the necessary information needed for the contextual representation, i.e., meaning, of a specific term. Other factors such as the context of usage (not shown here due to space limitations) can be combined with its contextual representation to restrict the search space. Hence, the user gets a complete picture regarding the semantic context of this and associated terms (see Figure 2) and is free to pick up a desired term(s) which would eventually lead him/her to candidate documents underlying the GCs and their meta-data schemas. Term entries in this GUI are mapped by means of the mapping services of a GC to the relevant schema terms found in the document meta-data schemas (in the same GC). This process involves IR cluster analysis techniques [4] to identify co–occurrence probabilities – representing the degree of similarity – in combination with term similarity and link similarity techniques and is described in some detail in in [10]. To provide the right ontological context for semantic term matching, we use again the lexicographic tool WordNet [9].

5 Schema Term Navigation and Querying

Information elicitation spans a spectrum of activities ranging from a search for a specific document term(s) (contained in possibly several documents) to a non-specific desire to understand what information is available in these documents and the nature of this information.

5.1 Navigation Techniques

There are two basic modes in which searching of the system may be organized. These search modes depend upon the nature of the information a user is attempting to access, and how this information relates to the document-base that user is operating from. In such cases the user is interested in finding out about a particular topic rather than a specific information (schema) item. We call this former form of exploration *index-driven*. Alternatively, if a user is seeking data which is closely related or allied to a particular document-base (currently under search), then searching may be organized around the weights of content links of this document-base to other GCs in the network. We refer to this form of exploration as *concept-driven*.

Index-driven navigation allows searchers to deal with a controlled amount of material at a time, while providing more detail as the user looks more closely and is related to the dynamic indexing schemes and incremental discovery of information. In order to traverse the index a user will have to decide on a number of key terms associated with a request for information. These terms can be selected from the terms shown in menus like those appearing in Figure 2. TopiCA locates the most general term, e.g., human-activity for the specified term, e.g., course, see Figure 3(a), and returns it to the user for verification. This process continues with more specific terms until the requested term (or its aliases) is located. A query graph structure is then generated on the basis of terms extracted by the user from the menu entries. This is compared against a context graph generated on the basis of the index structure in a GC. The query graph structure as well as the context graph structure of a GC are generated on the basis of terms extracted from WordNet entries which serves as a common ontology for comparisons. The comparison starts at the top of the index and gradually percolates down to the required level of specificity by following the terms at each level. In each step alternatives are proposed to the user (search engine). By matching the user query graph to its closest context graph we can obtain a number of DIRs (and documents) most closely associated to a search request.

Concept-driven searching is used when the user embarks on explorative searches and is most likely interested to find data closely related to a local document by following GC link-weights. We will use the GC connections shown in Figure 3(b) to illustrate this form of searching. The concept-driven search is based on the weights with which a specific document base, e.g., Accreditation, is linked to the various other GCs in the system. This document base's weight to the Education GC (its own GC) is 10/10, whereas its links to the Training, Employment, and Literature & Publications GCs are weighted with 8/10, 7/10 and 5/10, respectively. The Education GC is in closer proximity to the Accreditation document-base, followed by the Training, Employment and Publications GCs. The user may then chose to explore DIRs contained in the Education GC first. Subsequently, s/he may choose to explore the

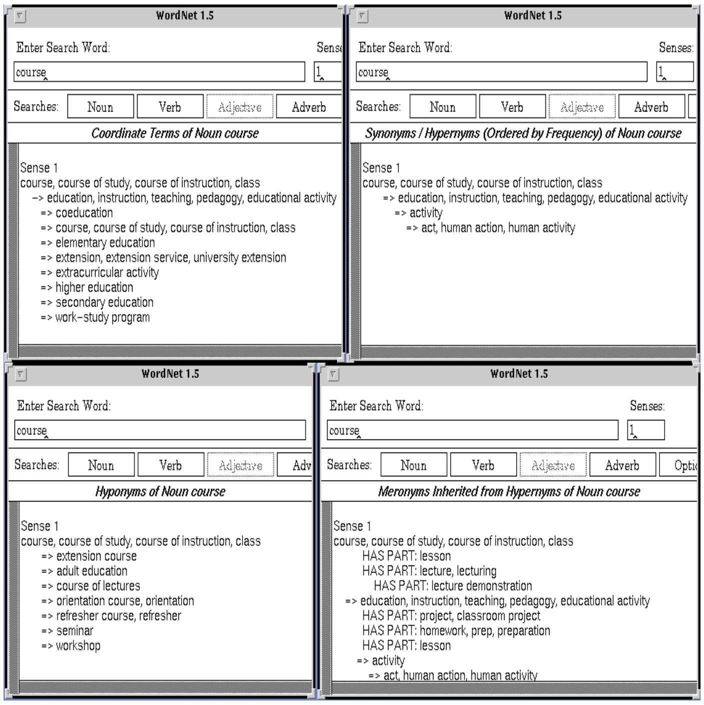

Figure 2: More contextual information regarding the term course.

Government Education Departments GC followed by the Publications GC and so on. The more weakly linked information is, the more general and the more ambiguous it tends to become. The two modes of navigation can be mixed: when exploring these GCs the user may embark on index-driven navigation to gain more insight into the concept found.

5.2 Querying of Domain Meta-Data

When the user needs to further explore the search target, *intensional*, or schema queries [13] – which return meta–data terms from selected schema terms – can be posed to further restrict the information space and clarify the meaning of the information items under exploration. Intensional queries are particularly useful for assisting users who are unfamiliar with the vocabulary of terms that can be used in connection with distributed searches on an FDL. Sample intensional queries related to the GCs in the previous sections may include the following:

query-1: *Give me all topics similar to accreditation under sense education OR government.*

query-2: *Give me all terms more specific than course and all their parts under sense education.*

query-3: *Give me all documents dealing with education under sense learning AND traffic under sense moving passengers.*

query-4: *Give me all documents similar to author = "S. Ceri" AND "P. Fraternali" AND title = "Designing Database Applications with Objects and Rules".*

Note that query-3 returns documents which belong to the intersection of two seemingly unrelated GCs (Education and Traffic), wile query-4 tries to match a certain book pattern (through its associated DIR) to that of other documents.

6 Summary

This paper described the fundamental aspects of a semantically oriented, scalable and configurable information infrastructure that supports interoperability across subject domains in federated digital libraries. The proposed logical architecture extracts semantics for documents and creates dynamic clusters of documents centered around common topics interest (viz the generic concepts). Large-scale searching is guided by a combination of lexical, structural and semantic aspects of document index records in order to reveal more meaning both about the contents of a requested information

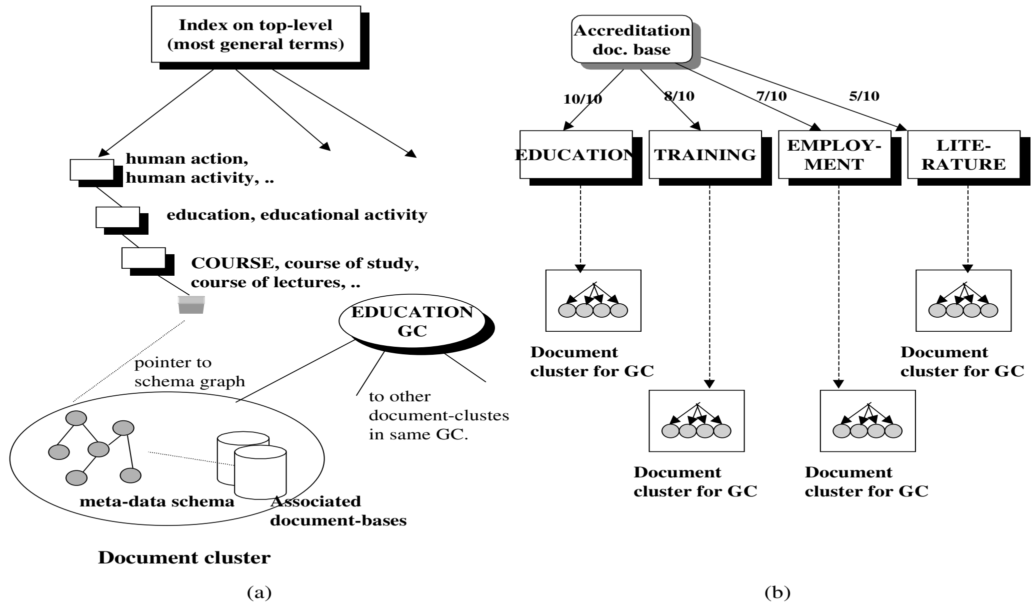

Figure 3: Navigation modes in TopiCA.

item and about its placement within a given document context. To surmount semantic-drifts and the terminology problem and enhance document retrieval, alternative search terms and terms senses are suggested to users. This architecture enables users to gather and rearrange information from multiple digital libraries in an intuitive and easily understandable manner.

An initial experimental prototype of TopiCA using the WordNet lexicographic substrate was implemented on Sun SparcStations under Solaris 2 using GNU C++ and CGI scripts. Parts of this system is now reimplemented in the context of the Decomate II Esprit project which aims to build the European Digital Library for Economics. The main aim of the project is to unify existing library systems of geographically distributed libraries, including the heterogeneous databases they own, under one single Web-based user interface.

References

[1] M. Agosti, M. Melucci, F. Crestani "Automatic Authoring and Construction of Hypermedia for Information Retrieval", *Multi-Media Systems*, **38**, 11, (1995).

[2] P. Atzeni, G. Mecca, P. Merialdo "To Weave the Web", *Procs 23rd VLDB Conf.*, Athens, Greece, Sept. 1997.

[3] M. Dunlop, C. van Rijsbergen "Hypermedia and Free Text Retrieval", *Information Processing and Management*, **29**, 3, (1993).

[4] Everitt B. "Cluster Analysis", *Heinemann Educational Books Ltd.*, Great Britain, (1981).

[5] "The InfoHarness Project" http://lsdis.cs.uga.edu/proj/proj.html.

[6] S. Jones et. al. "Interactive Thesaurus Navigation: Intelligence Rules", *Journal of the American Society for Information Science*, **46**, 1, (1995).

[7] Kirriemuir J. et al., "Cross-Searching Subject Gateways", *D-Lib Magazine*, January (1998).

[8] A. Levy, A. Rajaraman, J.J. Ordille "Querying Heterogeneous Information Sources using Source Descriptions", *Procs 22nd VLDB Conf.*, Bombay, India, Sept. 1996.

[9] Miller G. "WordNet: A Lexical Database for English", Communications of ACM, **38**, 11, Nov. (1995).

[10] Milliner S., Papazoglou M., Weigand H. "Linguistic Tool based Information Elicitation in Large Heterogeneous Database Networks", *NLDB '96 Natural Language and Databases Workshop*, Amsterdam, June (1996).

[11] "OMNI, Organizing Medical Networked Information", http://omni.ac.uk/

[12] A. Ouksel "A Framework for a Scalable Agent Architecture of Cooperating Knowledge Sources" in Intelligent Information Agents, M. Klusch (ed), Springer Verlag, to appear Feb. 1999.

[13] M. Papazoglou "Unraveling the Semantics of Conceptual Schemas", *Communications of ACM*, **38**, 9, Sept. (1995).

[14] M. Parunak "Don't Link Me In: Set Based Hypermedia for Taxonomic Reasoning", *Procs ACM Conference on Hypertext*, San Antonio, (1991).

[15] Schatz R.B., et. al "Interactive Term Suggestion for Users of Digital Libraries", *1st ACM International Conf. on Digital Libraries*, Bethesda MD, March (1996), 126–133.

[16] "SOSIG: The Social Science Information Gateway", http://www.sosig.ac.uk/

[17] T. Smith "The Meta-Data Information Environment of Digital Libraries", *Digital Libraries Magazine*, July/August 1996.

[18] P. Willett "Recent Trends in Hierarchical Document Clustering: A Critical Review", *Information Processing and Management*, **24**, 5, (1988).