13

TopiCA: A Semantic Framework for Landscaping the Information Space in Federated Digital Libraries

M.P. Papazoglou[†], *H. Weigand* [†], *S. Milliner* [‡]

[†] *Tilburg University*
INFOLAB
P.O. Box 90153,
5000 LE Tilburg,
The Netherlands
{*mikep, weigand*}*@kub.nl*

[‡] *Queensland University of Technology*
School of Information Systems
GPO Box 2434,
Brisbane QLD 4001,
Australia
steve@icis.qut.edu.au

Abstract

TopiCA is an architecture for a distributed information navigation infrastructure which aims to organize the information space across interconnected digital libraries around subject-areas to provide a sound basis for semantic information discovery and retrieval. TopiCA links document-bases by concentrating on topic similarity in such a way that clusters of document combinations are formed. This system provides topic-level browsing support and couples the search process with a lexicographic facility that makes use of a controlled vocabulary to find variant forms of terms and support term suggestion.

The paper describes the fundamentals of the TopiCA-topology and explains how this system imposes a semantic organization on the distributed information space to support a suite of activities ranging from a well-defined search for a specific document to a non-specific desire to understand what information is available in a federation of digital libraries.

Keywords

Digital libraries, meta-data, semantics, document classification, search engines, subject-based document discovery

1 INTRODUCTION

Despite the advances in database technology, users of on-line information are often overwhelmed by the amount of information on-line, the subject

Data Mining and Reverse Engineering S. Spaccapietra & F. Maryanski (Eds.)

knowledge required to access inter-networked information resources, as well as the constant influx of new information. This is especially true for *federated digital libraries* (FDLs). FDLs are forms of digital libraries (DLs) with spatial distribution whose aim is to make distributed collections of heterogeneous documents appear to be a single (virtually) integrated collection. In such federated digital libraries the difficulty lies in transforming a federation of multiple sources of documents into a single logical source. Individual digital libraries in an FDL are usually managed autonomously and have heterogeneity in database schemata, terminology, data formats and especially semantics. This heterogeneity causes difficulty to utilize multiple information sources effectively in a coherent manner.

The logical coherent accessing of information in FDLs is an involved process not only due to the sheer volume of information available, but also because of the terminology problem, which is consequence of the diversity of expertise and backgrounds of system designers and users. Conventional information retrieval technology matches terms specified by users to terms occuring in a digital collection. This term-matching is most effective when specialists access materials in their own area of expertise with precise terminology. Broadening access to a collection of documents in diverse digital libraries requires sophisticated search techniques to provide effective support to users working across very many areas of expertise. Specialists in even closely related subject areas cannot usually find relevant materials using current information systems technology. They know the concepts, but not the right terms. Searchers are thus presented with the problem of gaining adequate knowledge of a potentially huge dynamic system, in order to access and combine information across digital libraries in a coherent and logical manner.

Today the most popular information retrieval mechanisms for FDLs are provided by the prevailing Internet WWW-based software and are based either on either keyword searches (e.g., the Lycos server at CMU or the Yahoo server at Stanford) or hyper-text browsing (e.g., Netscape browser). Keyword searches result in relatively low precision and poor recall, due to the limitations of current indexing schemes, as well as the inability of searchers to fully articulate their needs. Current browsing on the Web consists of the traversal of (1) hyper-text-style links between explicitly related documents, and (2) indexes and meta-indexes, which are usually structured according to organization, and are almost always incomplete in their coverage (Francis *et al.* 1995). What is required is complete *topic-level* browsing in the context of FDLs. This implies that all document resources are linked, at a logical level, according to topic. Moreover, browsing should assist the users in the process of information discovery by allowing them to visualize relations among searchable terms, and by interactively providing the searcher with conceptual maps that offer alternative search terms. Interactive term suggestion, where the system suggests terms for the user to choose, can also significantly enhance retrieval effectiveness (Schatz *et al.* 1996b).

The general search problem in FDLs spans a spectrum of activities ranging from a well-defined search for a specific document to a non-specific desire to understand what information is available. The networked information environment introduces a number of new challenges for finding the right type of information in networks of digital libraries. This brings to the forefront the necessity of building new, more powerful, user workspaces for discovering and using information in this increasingly rich and varied environment. There is a great need for the development of seamless interfaces so that searchers can productively spend their time with the content of the information rather than trying to navigate huge dissimilar information spaces ranging across networked digital libraries. In order to improve the efficiency of searching in FDLs, the first requirement is to partition the FDL information space into distinct subject categories meaningful to searchers. Subject partitioning creates smaller index databases, which are more efficient for searching. In addition, a subject category created as a result of classification of information can also substantially aid searchers.

Our research work concentrates on developing a semantic framework, referred to as a *Topic*-based Document *C*lustering *A*rchitecture (*TopiCA*), to support searches across federated digital libraries relying on document content rather than their structure. In particular, our work focuses on extracting semantics from document indexing records across subject domains using the notion of concept spaces and proposes a logical organization of the search space in FDLs around subject-areas to provide a sound basis for semantic retrieval. Semantic links are created between document resources that are topically related. Indexes are searched by routing from topic to topic (versus document to document) until the appropriate documents are found. Browsing is accomplished in a similar fashion, except that the movement from topic to topic is user induced. Moreover, the search/browsing process is intertwined with linguistic support facilities that make use of a controlled vocabulary to solve problem relating to differing semantics and terminology deviations.

Our approach to FDLs places emphasis in assisting users to discover, understand and logically cross-correlate distributed information by supporting user exploration. This can be contrasted with conventional distributed/federated database approaches, e.g., TSIMMIS (Papakonstantinou *et al.* 1996), where emphasis lies in data integration and extraction of heterogeneous data by means of a common query language rather than on information finding. In the following sections, we contrast *TopiCA* with current Web practices in FDLs. Subsequently, we discuss the issue of meta-data representation in FDLs and propose a set of criteria for achieving semantic searches in FDLs and explain how *TopiCA* meets these criteria.

2 REVIEW OF CURRENT WEB PRACTICES IN FDLS

Related work can be classified into the following categories: web-based resource discovery, and document clustering.

2.1 *TopiCA* versus Database Search Engines

The use of the World Wide Web (WWW) has led to the development of a variety of search engines which attempt to locate a large number of WWW documents by indexing large portions of the Web: they recursively enumerate hyper-text links starting with some known documents. Currently, the most effective way of searching FDLs is through one of the growing number of available database search engines. These search engines are based on tools which index document collections and a directory (database) server that guides users towards the independent indexes and can be categorized into two types:

1. Those that attempt to index the entire Web (*centralized database approach*).
2. Those that attempt to index a selected portion of the Web (*federated database approach*).

Centralized Database search engines such as Lycos (Mauldin *et al.* 1994), Web Crawler (Pinkerton 1994), ALIWEB (Koster 1994) are manual indexing schemes that rely on search engines which "crawl" the network compiling a master index. The index can then be used as a basis for keyword searches. These systems are not scalable because they use a global indexing strategy, i.e., they attempt to build one central database that indexes everything. Such indexing schemes are rather primitive as they cannot focus their content on a specific topic (or categorize documents for that matter). Indexes are most useful in small collections within a given domain. As the scope of their coverage expands, indexes succumb to problems of large retrieval sets and problems of cross disciplinary semantic drift.

Some of the above limitations are addressed by *federated database search engines* such as Harvest (Bowman *et al.* 1995). The Harvest information discovery and access system (Bowman *et al.* 1995) provides an integrated set of tools for gathering information from diverse Internet servers and a customized view into what has been "harvested". It builds topic-specific content indexes (summaries from distributed information), provides efficient search mechanisms, and caches objects as they are retrieved across the Internet. Each local search engine builds a specialized directory for a certain domain of documents. Federated search engines scan those directories and form federated directories which aggregate documents according to application-specific needs. However, the Harvest Registry acts like a WAIS directory of servers: it only supports

the ability to click on brokers which provide the indexing and query facilities to the gathered information. This is due to the fact that the level of granularity is still a document (rather than an data object in the case of databases). Such directories play the same role as federated dictionaries in multi-database systems.

In contrast to *TopiCA*, none of the above systems attempts to logically structure the search space, partition and categorize information, or address terminological problems. Moreover, they do not provide any means of semantic support, e.g., term disambiguation, term suggestions, subject-driven searches, for user requests as suggested in this paper.

2.2 *TopiCA* versus Distributed Searching Techniques

There are many different distributed searching techniques. These range in functionality and complexity from strict traversal of a naming tree (Mokapetris 1995) to automatic dispersion of terms over a mesh of forward information servers (Centroids). One of the most popular approach was to build a static clustering of the entire collection of documents and then match a query to the cluster centroids (Willet 1988). Often a hierarchical clustering was used and a query was compared against each cluster in either a top-down or bottom-up manner. In most cases classification schemes were proposed to statically group the documents and reflect the kinds of queries that would be received, based on heavy reliance on subject codes in bibliographic search (Larson 1992).

Scatter/Gather (Hearst *et al.* 1996) is an approach that dynamically clusters small collections of documents for browsing large information spaces. It presents summaries of clusters to the user, who can then select a subset of these clusters for further consideration. The selected clusters are scattered into a small number of document groups and re-clustered on the fly. The new clusters then reveal their contents in more detail as the number of re-clustered documents is much smaller.

HyPursuit (Weiss *et al.* 1996) uses terms and hyper-links to cluster large collections of hyper-text documents that can be searched or combined into larger clusters. It also defines a framework for information retrieval services, such as query routing and refinement in a hierarchy of servers (storage for digital objects).

Most of the above techniques, though might be efficient for keyword-based searches, do not address the relevance of documents to specific query terms as expressed by the user.This is due to the fact that they assume user familiarity with the indexed keywords in the database and even the domain that they are searching.

TopiCA uses a different approach in that it defines broad subject areas for related document collections, generates term graphs for each index record based on a common linguistic tool, and then forms a concept space based on

pair-wise similarities of nodes in term graphs (again on basis of the common linguistic tool) and their link structure. Some similarity exists between our approach and that of HyPursuit, however, this system concentrates on hypertext document clustering mechanisms and does not address issues of topic-based navigation as described herein.

3 THE META-DATA ENVIRONMENT IN DIGITAL LIBRARIES

A traditional physical library is a single repository for materials from many sources which the user searches for information. A repository is just an organized collection in which documents and other information objects are indexed for effective search. A digital library is a group of these distributed repositories that the user perceives as a single logical repository (Schatz *et al.* 1996b). Surrogates of the documents in the library – called *document index records* (DIRs), or *meta-data* – are created for the purposes of value added by catalogers and indexers. Data are recorded in specific fields or subfields of these records, and are "finely searchable". This is a standard process with all libraries which employ highly skilled human indexers for such purposes.

3.1 Meta-Data in Traditional and Digital Libraries

Descriptive cataloging is probably the most important class of meta-data in traditional libraries. The well-established Anglo-American cataloging rules and the MARC interchange format (Library of Congress 1996) is the basis for virtually all existing library systems and has proven effective in creating and encoding descriptions of a great variety of content. However, the very complex rules require extensively trained catalogers for successful application.

The concept of meta-data (index records) when applied in the context of digital libraries 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 (Smith 1996). In contrast to traditional descriptive cataloging simpler descriptive rules are employed which are sufficiently simple to be understood and used by the wide range of authors and publishers who contribute information to the Internet. DIRs are used as document surrogates necessary for conducting search and retrieval. Like the actual documents it represents, the DIR contains full title and author information, a list of keywords, as well as the abstract as it appears in the document. Databases of DIRs are usually made of these descriptive elements and invert several of their data elements by creating an index (pointer) to the document file itself. In this paper we assume that the minimal service we would expect of a digital library is this form of indexing.

As an example of its use in the context of DLs the term meta-data has been used to describe the information of the "Dublin Core" (Weibel *et al.* 1996)

and the associated "Warwick Framework" (Lagoze 1996) which are intended to facilitate access to information available via the World Wide Web. The Core specifies the concrete syntax for a minimal set of descriptive elements that facilitate the description and the automated indexing of document-like networked objects, and the framework specifies a container architecture for aggregating additional meta-data objects for interchange. The elements of the Core include familiar descriptive data such as *author*, *title*, *subject*, *form*, *type*, *relationship* to other information objects, and so on. However, the descriptive rules suggested by the Core do not offer the retrieval precision, classification and organization that characterizes library cataloging. The Core is simply a method for abbreviated descriptive cataloging that permits untrained authors or editorial stuff to describe their digital (document) resources themselves.

3.2 Meta-Data in Federated Digital Libraries

It is difficult to support the federation of multiple DL resources into a single logical resource. Part of the difficulty lies in handling the documents which have differing structures, styles and terminology. Handling searches is also difficult. They must support different classification schemes so that document sources can be indexed in various ways at different levels of detail. Accordingly, in order to provide a coherent view of digital objects in an FDL, these must be described in a consistent fashion which can facilitate pro-active distributed document searching and retrieval so that the FDL is not merely a passive warehouse of navigatable information. To achieve this objective, FDL applications need to rely on a higher-level context of meta-data than descriptive cataloging to facilitate dealing with the problems of large-scale searches and cross disciplinary semantic drifts. This context should define how distinct sets of index-records can be aggregated logically to provide greater interoperability by allowing tools and searchers to selectively access individual document aggregations while ignoring others. This implies mapping both the form and the contents of document index records across subject domains.

To explicate the usefulness of meta-data in FDLs we outline four value-added meta-data functions that help enhance the quality of semantic searches in FDLs. These are briefly explored in terms of their role in an FDL in the following:

1. *A classification scheme:* should be used to group semantically related index records according to topicality, i.e., the subject of each resource, and a structure should be created to indicate relationships between document clusters. The general idea is to index document index records together into topically-coherent groups, and present textual summaries and a common structured vocabulary of topical terms to searchers for interaction. We refer to this framework which contains topical synoptic knowledge, regarding

semantically related collections of document index records, and a standard vocabulary for term suggestions as the *concept space* for a document cluster. The concept space provides the means for both broad categorization to identify relevant topics in digital libraries and specific categorization to select document resources within these libraries.

2. *Dynamic indexing schemes:* The requirement for a flexible classification scheme is based on the existence of flexible indexing schemes. To be effective search techniques must connect local topic-specific indexes and other pieces of information via a distributed multi-level indexing scheme. For instance, several local indexes can be combined to form a two-level index, in which the top level can only filter queries to a subset of the local indexes. The main advantages of such a scheme is that the index can be partitioned in several ways, which makes it scalable and the number of search terms is quite limited no matter how much data exists. In addition, indexing must be updatable to cater for fluctuations in the system (which will be more dynamic the larger it is), and allow for incremental user updates over time.
3. *Incremental discovery of information:* As users are confronted with a large, flat, disorganized information space it is only natural to support them in negotiating this space. Accordingly an information elicitation system should provide facilities to landscape the information available and allow the users to deal with a controlled amount of material at a time, while providing more detail as the user looks more closely.
4. *Domain specific query formulation assistance:* An important service is user assistance with the formulation of information retrieval queries. For example, users may not know or understand the idiosyncratic vocabularies used by information sources to describe their information artifacts and may not know how to relate their functional objectives to these descriptions. Any system that provides global information access must help the user formulate queries that will return more useful results and avoid inundating them with unwanted material. This can be achieved by allowing a query-based form of progressive discovery in which the user finds out about subject-areas of interest rather than specific information items (see previous item).

An analysis of the preceding requirements reveals several challenges relating to the mechanisms by which users locate, select and retrieve information resources, viz. document collections, in FDLs. One can therefore characterize the meta-data environment of an FDL in terms of a model involving a set of *services* for:

- Organizing the information space across federated collections of documents in DLs and providing serendipity, exploration and contextualization support so that users can achieve logical connections between known concepts and new terms. Also, coordinating user interactions with the meta-data environment.

- Resolving the terminology problem and "semantic drifts" which arise due to the terms that are indexed and are used differently across documents and index records in different DLs.
- Creating models of information items in a DL and providing access to these models for querying. Meta-data should be conveyed at a level above that of descriptive cataloging or emerging meta-data standards for digital libraries (e.g., the Dublin Core).
- Describing widely varying levels of aggregated information sources combined with the ability to describe the granularity of the components that comprise them.
- Constructing models of the user's query and user's workspace requirements.
- Making matches between the model of the user queries and models of information items in a DL.

In the following we describe a semantic framework that addresses several of these considerations in the context of multiple co-existing heterogeneous document collections in a federated DL environment.

4 THE *TOPICA* INFRASTRUCTURE

In FDLs users are confronted with a large, flat, disorganized information space and require support in negotiating this vast space. *TopiCA* is a logical distributed framework that achieves a semantic organization of the information space in DLs, according to topic-areas, and provides facilities to contextualize and landscape the information available. *TopiCA* allows a searcher to deal with a controlled amount of material at a time, while providing more detail as the searcher looks more closely.

To address these requirements *TopiCA* is based on semantic information embedded in DIRs to achieve clustering of related documents around *Global Concepts* (GCs) which materialize the concept space for a specific group of documents according to document with topicality (Papazoglou *et al.* 1996). A GC is a form of a logical object whose purpose is to cross-correlate, collate, and summarize the meta-data descriptions of semantically related network-accessible data. GCs achieve explicit semantic clustering between relevant pieces of information as they pertain to (and abstractly describe) areas of interest common to several DIRs. Thus, a contextualized FDL information space consists of a number of centroids of the inter–DIR information space, viz. the GCs, around which documents effectively cluster.

This section discusses how *TopiCA* provides a logical framework for forming multiple co-existing document cluster hierarchies and explains how topic-based searches are achieved.

4.1 Topic-based Information Spaces for Semantic Navigation

TopiCA can be viewed as a Web-space that encompasses collections of DIRs in networked digital libraries. *TopiCA* links, see Figure 1(a), denote general topic similarity between diverse DIRs. *TopiCA* clusters documents into topically-coherent groups, and presents descriptive term summaries and an extended vocabulary of terms for searching and querying a vastly distributed information space. Individual document cluster hierarchies are useful for browsing and searching large document collections scattered across discrete DLs because they organize the information space. The *TopiCA*'s clustering mechanism results in grouping DIR elements from diverse documents that share important common properties onto a generic concept, associating these properties with the GC representation, and regarding the GC as an atomic unit.

To put the organization of an concept space into perspective, we consider the case of the Accreditation document-base which contains collections of documents providing information about accreditation of courses and cross-institutional subjects, various private/public educational training information and other similar or related documents. In its original form the Accreditation document-base, maintains information only on education service providers, their courses, accreditation committee members, accreditation processes and related information. Figure 1 shows the Accreditation document-base along with a partial representation of its associated DIR. It also illustrates how this document-base (represented as oval) may become part of a larger DIR network by establishing weighted links to GCs implementing related areas of interest. Consequently, the Accreditation DIRs are not only able to source appropriate information from remote documents based on the same topic but also to provide *matching* information about enrollment programs, training schemes, research activities and publication data.

Navigation in *TopiCA* can be considered as browsing through DIRs exclusively at a "macro"-level, i.e., from topic area to topic area such as from educational training, to publications, government departments and so on. These topic areas are composed of a group of related DIRs and their underlying documents, see Figure 1(b). In addition, the topic-areas are interconnected by weighted links to make the searches more directed and meaningful, see Figure 1(a). This should be contrasted to the navigation style of web-based documents where navigation takes place at the "micro", viz. HTML, level and tends to be from individual document to individual document (Francis *et al.* 1995).

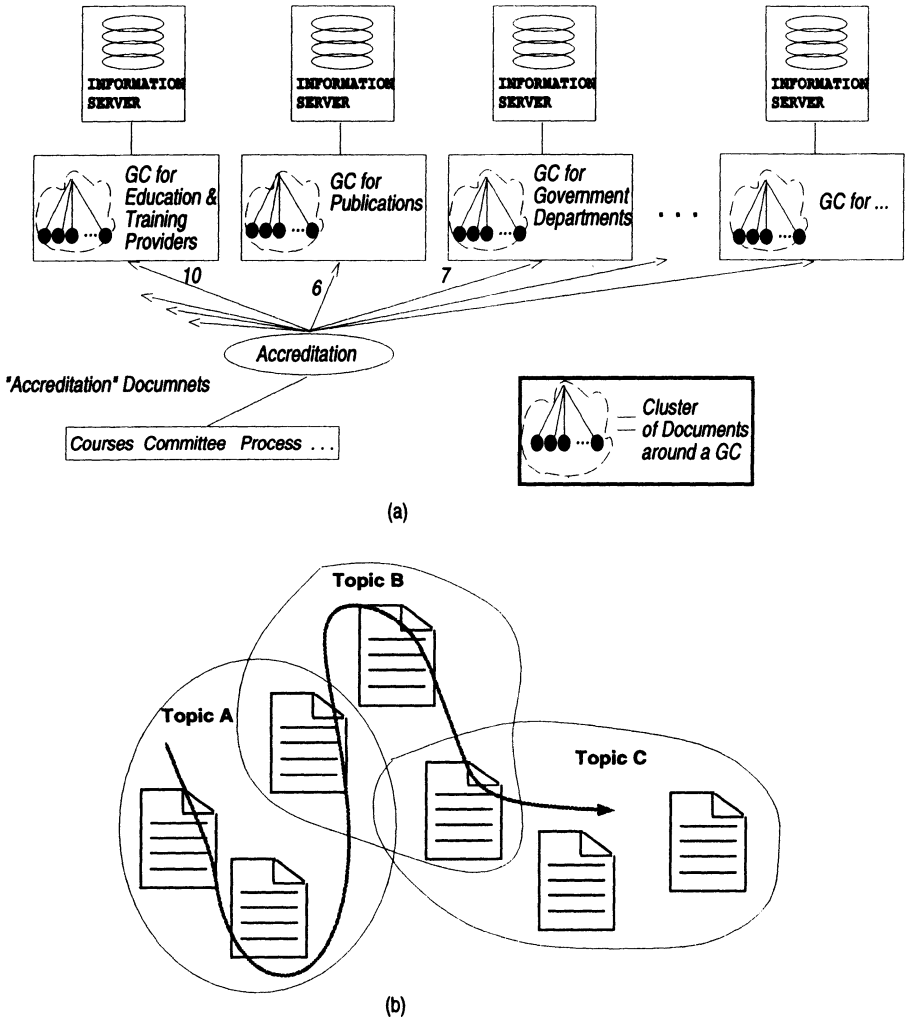

Figure 1 Connecting document bases.

4.2 Basic Components of the *TopiCA* Infrastructure

As illustrated in Figure 2, the basic component of the *TopiCA* infrastructure is a Meta-Data Schema (MDS) essentially describing each document DIR content plus some additional semantic information. The meta-data schema is essentially a summary of the resource it represents together with a standard vocabulary for handling queries directed to its underlying documents. Meta-data schemas of semantically related documents are installed in an information server, see Figure 2. The information server essentially adds the term vocabulary and the meta-data schemas of its underlying cluster of doc-

uments (which cluster around a specific GC) to its associated database. The meta-data schemas in an information server are then linked with selected sets of similar other meta-data schemas, viz. GCs, in other information servers, see gray lines in Figure 2. These links correspond to the weighted links in Figure 1(a) which interconnect semantically related GCs. Thus, a GC essentially becomes part of a node in a mesh network of meta-data schema collections.

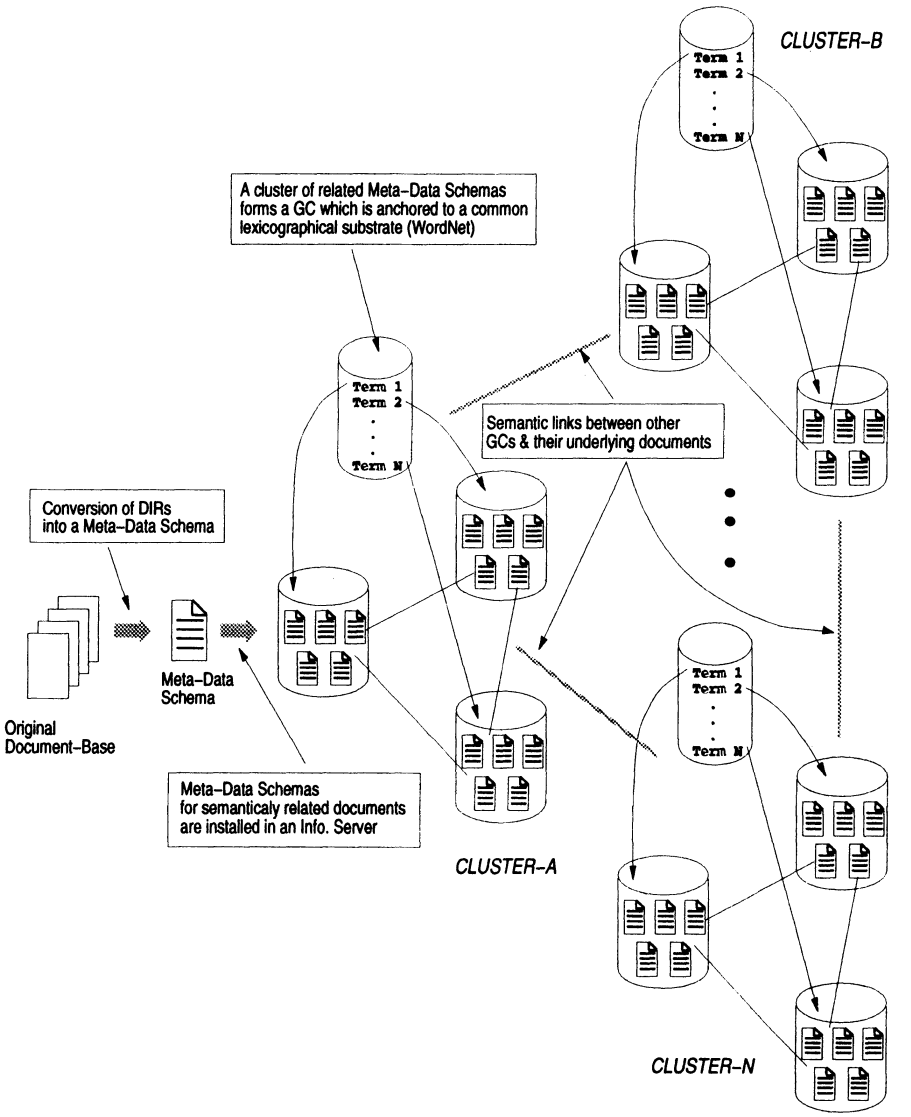

Figure 2 Components of the *TopiCA* infrastructure.

To facilitate clustering and discovery of information, we require that a DIR (and its underlying document, e.g., Accreditation) can be totally described in terms of three elements. These contain a synoptic description of the meta-data content of the document; associations between meta-data terms in the form of a semantic-net; and finally, links from these descriptions to other related document clusters in the network. Figure 3 illustrates the required elements of an meta-data schema: (1) a *feature descriptions*, (2) a *context graph*, and (3) a *GC connections* section. The feature descriptions section contains information about terms, composition of terms, remarks about the meaning of terms, and lists of keywords. This section also includes certain details such as: geographical location, access authorization and usage roles, explanations regarding term usage and definitions, domains of applicability and so on. The feature descriptions entries are partially generated by the lexicographic substrate that underlies a meta-data schema (see bottom left of Figure 3). This lexicographic tool is based on WordNet (Miller 1995) and supports semantic term matching through the use of an extensive semantic network of word meanings of terms connected by a variety of textual and semantic relations. WordNet places emphasis on word (term) structure (lexical relations) and semantics (word meanings). This comes in contrast to other ontologies such as Cyc (Cycorp 1995) which provides taxonomic classification of general concepts of human consensus reality and places less emphasis on word structure and word meanings. WordNet provides a variety of semantic relations, such as synonymy, antonymy (opposite-names), hyponymy (sub-names), meronymy (part-names), troponymy (manner-names), entailment and an extensive list of keywords to describe a term. More importantly, it allows users to choose between alternative senses of a polysemous term in order to distinguish between different sets of linguistic DL contexts in which the term can be used to express the term meaning. The context graph section contains a non–directed graph which connects term synopses (in the form of *term descriptor nodes*) found in the Accreditation DIR. These term descriptor nodes and their link structure are used in the clustering of documents to form the generic concepts. Each of the term descriptor nodes defines (in conjunction with its respective entry in the feature descriptions window) a common structured vocabulary of terms – describing the term in question, e.g., course, – and a specification of term relationships within that particular subject. To demonstrate this consider the previous example of the Accreditation document-base, which deals with academic institutions and accreditation processes. The DIR of this document-base contains terms such as courses, committees, (accreditation) processes, etc. The context graph created for this DIR (Figure 3) contains nodes which correspond to the terms committee, institutions, courses etc., while the context graph edges depict non-"colored" inter–connections (association, generalization, specialization or containment) between these terms. Finally, the GC connection section shows how the Accreditation DIR is related by means of link weights, to other GCs in the network.

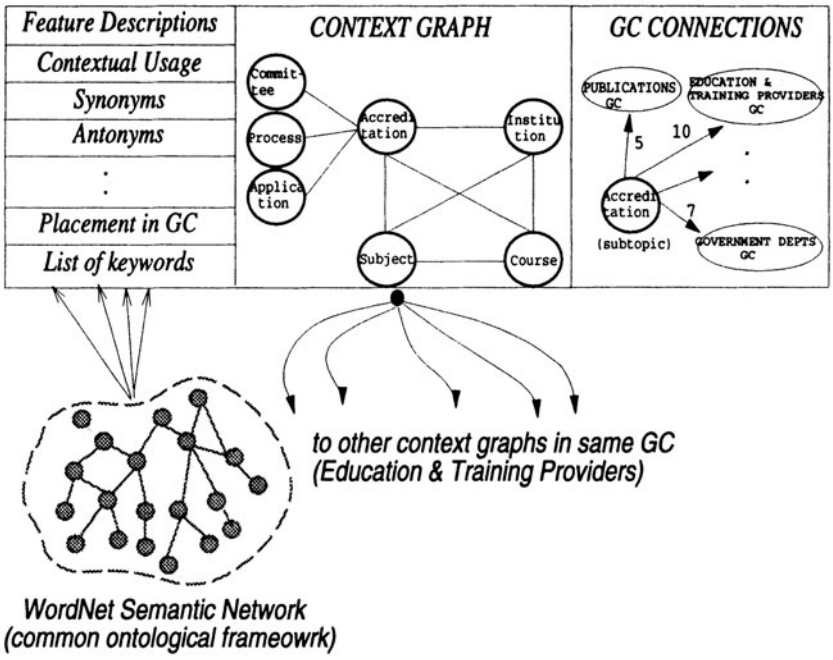

Figure 3 Meta-data schema of a document.

Consider the case of a searcher interested in documents related to the term "course" in the context of academic settings. This searcher may select the "Education & Training" GC from a menu containing all GCs in the network. Then s/he will be presented with all the meta-data terms contained in this GC, viz. its underlying meta-data schemas as illustrated in Figure 2. In our example, this GC contains three meta-data schemas highly related to it. These are: Accreditation, Education_and_Training and Enrollment_Program. Course is a term contained in the context graph node of Accreditation, depicted in Figure 3, and is clicked by the user to start the search process. The user is then prompted to answer whether s/he wishes to have more information about the context of this term. The *TopiCA* GUI output for the term descriptor "course" that a searcher has specified as his/her search target is depicted in Figure 4. This term is shown to have eight senses, but when the domain of discourse is limited to academic courses, then only one of the eight can occur. Once the user selects the appropriate sense a series of windows open to reveal the *topical context* (the vocabulary used to discuss a well-defined topic) of this particular term. Term entries in this GUI also point to other documents (via their respective DIRs in the same GC) that contain these and related

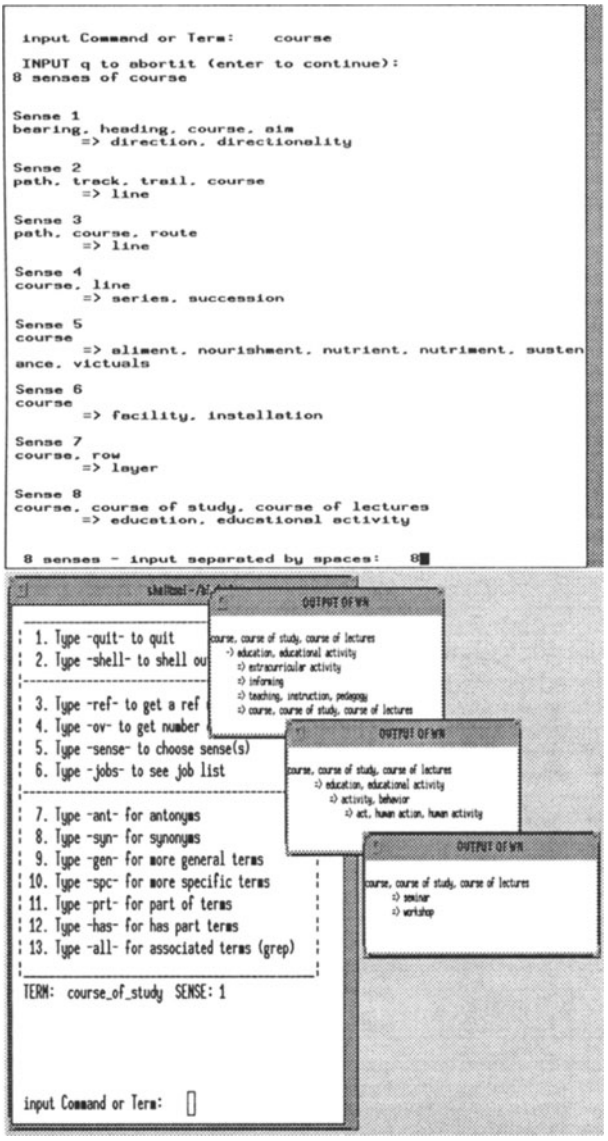

Figure 4 Semantic network output.

terms. More information about the different types of search and navigation are presented in section-6.

4.3 The *TopiCA* Logical Framework

Individual GCs are useful for browsing and searching large document collections because they organize the information space. The GC structure is akin to an associative thesaurus and on-line lexicon (created automatically for each subject category). Thesaurus-assisted explanations are created for each subject-based abstraction (GC-based information subspace) and serve as a means of disambiguating term meanings, and addressing terminology and semantic problems. A GC, thus, provides an *ontology* to describe a cross-document subject domain and facilitates access to the information by providing resources underlying this specific subject domain. For example, the Education and Training Providers ontology, which is partially materialized by WordNet, provides a common terminology basis upon which documents dealing with enrollments, courses, training, accreditation, etc. (Figure 1(a)), achieve implicitly knowledge of each others information content. The advantage of forming such document clusters, is that searches are goal–driven* and the number of potential inter–DL interactions is restricted substantially.

Currently, in *TopiCA* human indexers assign "aboutness" to DIRs and documents by means of large document collection and GC labels, e.g., Accreditation or Education_and_Training. Human indexers also decide the degree of relatedness between DIRs and GCs. By *strongly linking* to a certain GC, e.g., with a factor 10/10, DIR nodes agree to associate with each other and thus inter–node organization is achieved implicitly. Each of these DIR nodes may also link less strongly to other GCs which have their own associated cluster of DIR documents, see Figure 1(a). A single DIR collection, e.g., Accreditation, may be simultaneously involved in several clusters of DIRs (information sub-spaces) to varying degrees, as dictated by the weights of its links to the various GCs. The resulting GC structure forms a massive dynamic network, resembling a cluster–based *associative network* (Findler 1979) (a variant of semantic networks that uses numerically weighted similarity links). This type of content-based clustering of the searchable information space provides convenient abstraction demarcators for both the searchers and the system to make their searches more targeted and effective.

Overall an FDL may be viewed in terms of a logical hierarchy, see Figure 5. Leaf nodes in this hierarchy are at the bottom-level of the hierarchy and represent documents of the information space in FDLs. Leaf nodes within this hierarchy are single documents and interior nodes correspond to single DIRs describing these documents while the topmost nodes represent clusters of documents. The middle level represents a simplified view of the *meta–data schema* for related document collections in an object-oriented form. The meta-data-level indexes and returns pointers to leaf documents that reside on DL sites. The top most level corresponds to the concept space (GC) level.

*A goal–driven search accepts a high–level request indicating what a user requires and is responsible for deciding where and how to satisfy it.

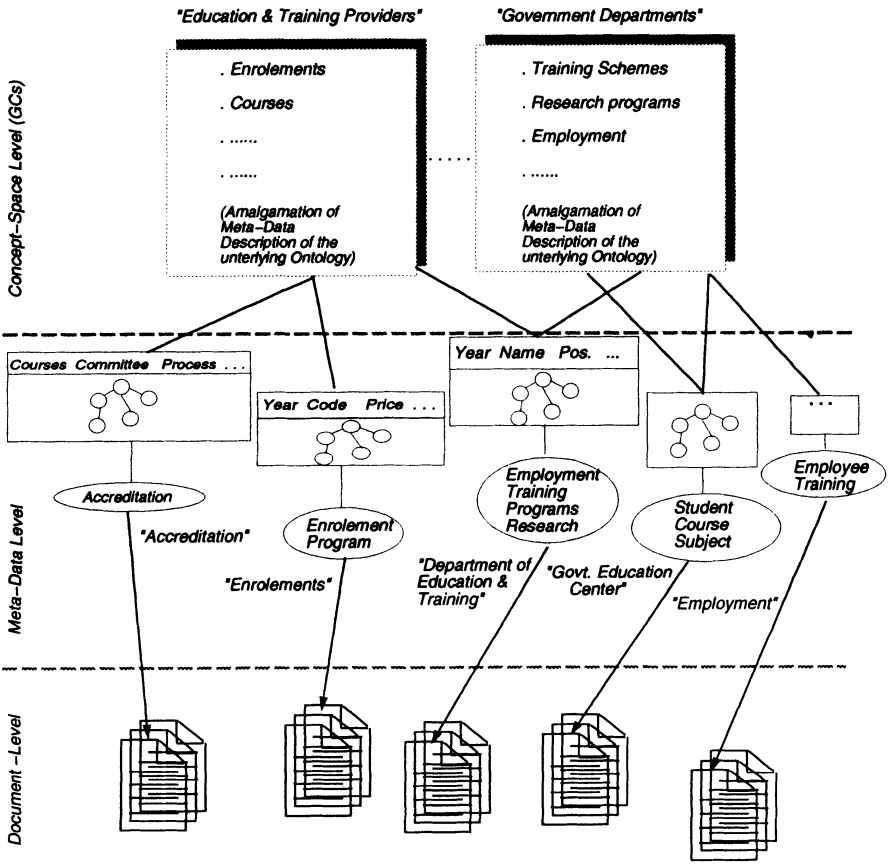

Figure 5 Levels of abstraction in the tiered inter-document organization.

This level contains abstract dynamic objects which implement the clustering of related DIRs and materialize the GC portions in an object-oriented form. This level corresponds to meta-data schemas such as the one presented in Figure 3. It is expected that there will be a restricted number of GCs per domain.

This tiered architecture is the key ingredient to information discovery in FDLs. It provides the ability to describe varying levels of aggregated document sources and 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, to its description in the meta-data level, and finally to the concept space level where the entire semantic context – as well as patterns of usage – of a term can be found). In this way, we can construct meaningful and scalable information partitions, which classify network-wide available information, and provide users with the means of coherent access to the body of information contained in an FDL conceptual space. User queries about document terms operate always in a top-down fashion, i.e., from the most

general (i.e., GC-level such as "Education & Training") to the most specific (i.e., document-level) level of detail and can be expanded to include a set of additional terms which do not match the query but which are semantically related to query items.

5 FORMING THE SEARCH SPACE

In the following we describe a general methodology that aids in clustering DIRs and creating their corresponding generic concepts. Key criteria that have guided this methodology are: scalability, design simplicity and easy to use structuring mechanisms based on object–orientation.

5.1 Similarity-based Clustering

Similarity-based clustering of documents organizes document index records into related groups based on the terms (term descriptor nodes) they contain and the link structure of their context graphs. Our method for comparing assumes that both document index records – to be compared – are represented as context graphs whose term nodes are grounded on the massive WordNet thesaurus.

Our clustering algorithm determines the similarity between two graphs (representing two different document meta-data) in two phases. Firstly, a pairwise-similarity of nodes in two context graphs is computed. From this an initial "pairing" of the nodes is determined. In the second step a comparison of the link structure of two context graphs is made based on the inter–node pairings and a semantic distance value is calculated. Hence, we consider localized equivalence of graph nodes, with respect to the global structures of their context graphs. This essentially corresponds to a grouping of documents based on their specialized subject areas.

Term-based Similarity: this is calculated using cluster analysis techniques (Everitt 1981) to identify co–occurrence probabilities – representing the degree of similarity – between two discrete terms. Our similarity metric is based on the meaning of the collection of terms representing the topical context (viz. semantic-levels) of a particular term, e.g., course, and the synonyms of these, see Figure 4. The comparison is based on: a conversion of each context graph node (e.g., term descriptor) Committee, Process, Subject, Course, etc. (see Figure 3) to a corresponding matrix of noun terms (containing the entire topical context of a term); and a subsequent comparison of terms within these matrixes.

A matrix $a_{n,m}$ of (noun) terms, representing the topical context of a particular term, $a_{i,1}$ (course say), will correspond to the name of the term

descriptor in the context graph. The synonyms of this term will be $a_{i,2}$, $a_{i,3}$ ... $a_{i,m}$ (course-of-study, course-of-lectures). Terms $a_{i-x,j}$ $(x > 0)$, e.g., education, educational-activity, will be more general than terms $a_{i,j}$, while terms $a_{i+x,j}$ will be more specific, e.g., CS-course. In the final step, all synonyms for these terms are generated to produce the node's a complete *topical description matrix* $a_{n,m}$ for a specific term.

Similarity analysis is mainly based on statistical co–occurrences of term descriptor objects based on techniques which has been successfully used for automatic thesaurus generation of textual databases (Chen 1994). To provide the right ontological context for semantic term matching, we use the massive semantic net WordNet (Miller 1995).

Comparison of the conceptual structure of two context graphs: when pairwise similarities between all relevant term pairs have been established, a hierarchical agglomerative cluster generation process can be adopted (Salton 1989). To determine the structural and semantic similarity between two graphs, we based our algorithms regarding conceptual similarity between terms on heuristics–guided spreading activation algorithms, and on work in the information retrieval (IR) area presented in (Rada 1989), (Kim *et al.* 1990). These approaches take advantage of the semantics in a hierarchical thesaurus representing relationships between index terms. The algorithms calculate the conceptual closeness between two index terms, interpreting the conceptual distance between two terms as the topological distance of the two terms in the hierarchical thesaurus. Accordingly, semantic relatedness is based on an aggregate of the interconnections between two nodes (representing terms) in a context graph. During this process similarity between nodes (term descriptors) is established by considering the edges separating the nodes in the context graph as well as the actual graph structure. The comparison and clustering process is described in (Milliner *et al.* 1996).

Once similarity between nodes has been established context graphs are aggregated (on a per topic basis) to create GCs. The aggregation of the context graphs (Figure 3) from various database nodes, results in the clustering of inter–related database schemas. For each database node group, a GC is created to represent the area of interest (or concept) that the group embodies, e.g., Education and Training Providers GC for the Employee Training, Accreditation, and Government Education Center databases as depicted in Figure 5.

6 NAVIGATING THE CONTEXTUAL SPACE

Keyword-searching in FDLs is only one part of the overall searching process. The search process should also provide *relevance feedback* and *vocabulary switching*, i.e., alternative term matching. Relevance feedback is an important part of any searching process as the searcher indicates to the search system

which (keyword) matching resources are in fact good matches. The system then finds document sources related to the good matches. Another useful function of a searching system is to suggest alternative terms that the searcher can may use in his/her search efforts (as we already explained in sections 1 and 4.2). Because related document sources are logically interconnected and based on a common lexicographic framework both these types of features can be efficiently supported in *TopiCA*. For the case of relevance feedback, a *TopiCA* search process can efficiently traverse the links surrounding a DIR returning to the user those meta-data schemas that have the most terms in common with the selected meta-data schema. For the case of suggesting alternative terms, a search process can refer to the WordNet entries underlying a meta-data schema and suggest alternatives and then link those to their related meta-data schemas. Note that in neither of the previous two cases is the searcher explicitly aware of the actual *TopiCA* links. The navigation process traverses semantically related links and returns the results to the searcher.

There are two basic modes in which searching of *TopiCA* may be organized. These search modes depend upon the nature of the information a searcher is attempting to access. Serendipity, exploration and contextualization are supported by means of indexing based upon terms contained in the DIR context graphs. In such cases the user is interested in finding out about a particular subject-area rather than a specific information item. Alternatively, if a user may seek data which is closely related or allied to a particular local document, then searching may be organized around the weights of links to the GCs. We call the former form of exploration *index-driven* while we refer to the later 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 browse topic-specific information, a hierarchical tree like index structure is constructed based on the merging of DIR context graph nodes. The first level of the index tree comprises a number of tree-nodes each representing the merger of several context graph term nodes based on the first row term equivalence(s) in their respective topical description matrixes (i.e., their most general terms). The second level of the tree index is constructed by considering equivalences between the first two rows of each description graph node's topical description matrix, i.e., more specific terms. This process of considering more and more matrix rows continues until the bottom most level of the tree is arrived at - this level contains the entire collection of all descriptions nodes with pointers to their description graph of origin, see Figure 6.

In order to traverse the index a user will have to decide on a number of key terms associated with a request for information, and then select synonyms or more general (and perhaps more specific) derivatives of these key terms. The resulting query structure - generated on the basis of terms extracted by the

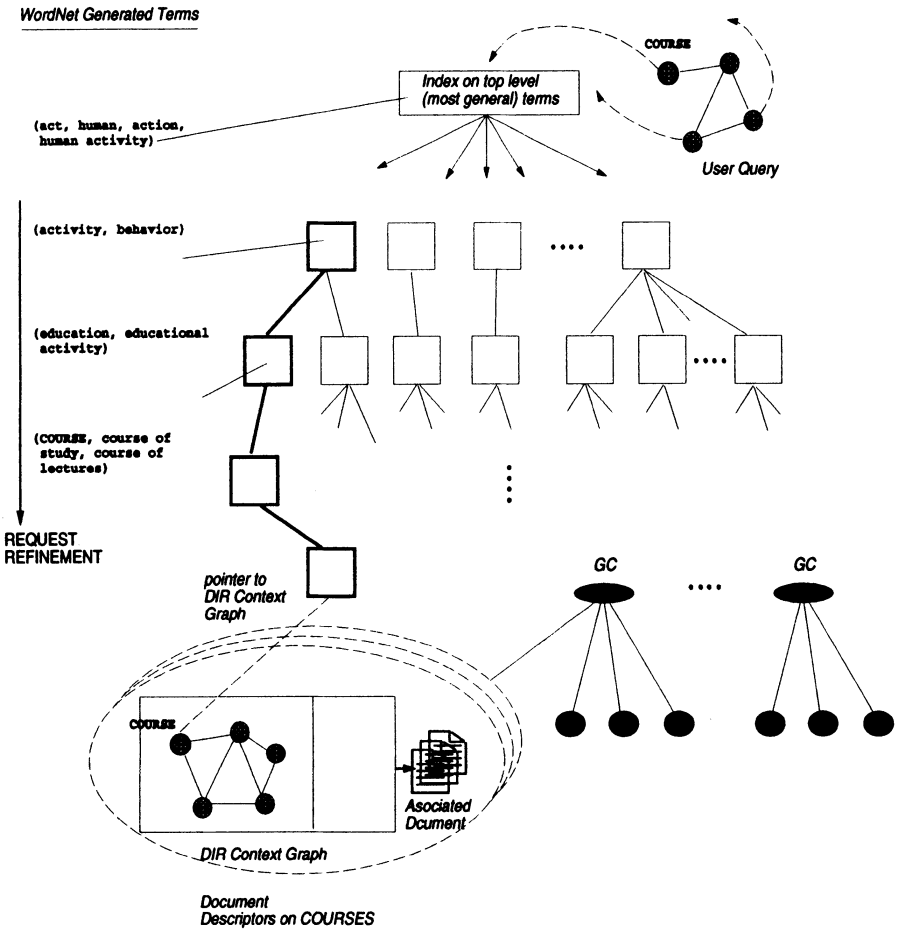

Figure 6 Index-driven exploration of terms in FDLs.

user from WordNet entries - can then be compared against the index structure in a GC. 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. Figure 6 depicts this process in terms of a user query requesting information about courses at various institutions. The user's description graph of the query contains a node Course and this concept is used to traverse the index and arrive at the collection of documents which include this term in their own descriptions. The index-driven navigation process starts with the most general terms, e.g., activity, which are presented to the user. These terms are generated by WordNet and are presented to the user for selection. Once the user has selected a general term, most specific terms are revealed, e.g., education. Once the user selects the term that matches his/her query then a link is established with the context graphs of all documents containing the desired

term, e.g., course. In this way the user can obtain contextual information and possibly an abstract regarding potentially matching documents and then s/he can decide whether a proposed document is useful or not. This hierarchical form of navigation guarantees that a user supplied term correlates semantically with the content of a document. The process is then repeated for all the other terms in the user's query graph (i.e. the remaining unlabeled nodes in Figure 6). By matching the user query graph nodes to description nodes, we can infer a number of DIRs (and documents) most closely associated to the user query.

Concept-driven searching is used when the user is seeking to find data closely related to a local document by following GC link-weights. We will use the GC connections section in Figure 3 to illustrate this form of searching. The concept-driven search is based on the weights with which the Accreditation document-base is linked to the various other GCs in the system. This node's weight to the Education and Training Providers GC (its own GC) is 10/10, whereas its links to the Government Education Departments and Publications GCs are weighted with 7/10 and 5/10, respectively. The Education and Training Providers GC is in closer proximity to the Accreditation node, followed by the Government Education Departments and Publications GCs. The user may then chose to explore DIRs contained in the Education and Training Providers GC first. Subsequently, s/he may choose to explore the Government Education Departments GC followed by the Publications GC. The more weakly linked information is, the more general and the more ambiguous it tends to become. When exploring these GCs the user may embark on index-driven navigation, as described above.

When the number of terms displayed after a concept-driven search is low enough and includes a sketchy description of the search target, *intentional* queries (Papazoglou 1995) – which return meta–data from selected DIRs – can be posed to further restrict the information space and clarify the meaning of the displayed information items. The intentional querying pane allows the user to ask for properties and documentation related to the conceptual structure of meta-data schema. It also allows the merging of intentional queries with traditional content–based searches, and supports incremental refinement of the information discovery process. This iterative process prevents zero–hit queries and supports the formulation of meaningful multi-document requests.

TopiCA can also handle other unconventional forms of querying, such as *querying by example.* For example, it is possible to get responses for queries like the following:

Give all documents similar to author = "S. Ceri" and "P. Fraternali" and tite = "Designing Database Applications with Objects and Rules".

The keyword "similar" is a predefined function which may compute all terms related with a weight of 8/10 or above. This query tries to match a book pattern to that of other documents.

7 *TOPICA* AND MULTI-LINGUAL DIGITAL LIBRARIES

Over the past few years we have experienced an enormous growth of interest in the construction of virtual DLs of world-wide distributed document-bases based not only on English but also other natural languages. Thus the question of multi-lingual access and multi-lingual information discovery and retrieval is becoming increasingly relevant. In this section we briefly discuss some of the issues related to this problem and describe our own approach to it. More specifically we we describe the architecture of a Lexicon Management System (LMS) to support the *TopiCA* framework for multi-lingual DLs. This LMS is currently being developed as part of the ESPRIT project TREVI (Weigand 1997) which supports news filtering and enrichment based on user profiles. In the following we will restrict our attention to the LMS and explain how it can be combined with the *TopiCA* environment and used in multi-lingual FDLs.

The LMS offers a concept space similar to what has been described before. User subscribers (or their mediators) can browse through this concept space to narrow down their profile. Hence, a user profile does not consist of words at string level, but of concept references. At the input side of the LMS, a parser and subject identification module scan incoming news messages disambiguating terms and in this way come up with a DIR in the form of concept references which collectively describe a news clip. User profiles and DIRs are matched taking the semantic relations between concepts into account and can join the *TopiCA* framework and thus become part of a larger document network of related materials.

The LMS consists of two main classes: the class of lexicals and the class of concepts. The relationship between the two can be expressed as: *"a lexical L can be used to evoke concept C"*. Grammatical information, such as inflection rules, are attached to the lexical, whereas semantic links such as hyponymy are made between concepts. Synonym sets are not explicit, but defined as the set of lexicals connected to a certain concept.

One advantage of the separation of lexicals and concepts is that the LMS can support different languages with a single concept space (see (Carbonell *et al.* 1995)). The LMS represents basic semantic relations between words for English and Spanish taking WordNet as its starting point. For each of the languages involved mono-lingual word-networks are created maintaining language specific nuances. Both word-networks share a common core-concept and multi-lingual relations will be mapped from each individual word-network to a structure based on standard WordNet meanings. Such relations form the basis of an inter-lingual index. Although it is possible that a certain concept has an expression in one language and not in another, or that subtle nuance differences may exist, most core-concepts are essentially identical. This work presents many similarities with the general purpose multi-lingual ontology being developed as part of the EuroWordNet project (Gilarranz *et al.* 1997).

As an example, let us assume that a user has expressed in his/her profile interest in Israeli politics and consider the following news item:

'Netanyahu vows to battle on after escaping charges
By Anton La Guardia in Jerusalem

ISRAEL'S Prime Minister, Benjamin Netanyahu,
was struggling to hold his coalition together last
night although it was decided yesterday not to
charge him with fraud and breach of trust.'

The LMS parser extracts at least the following terms:

Netanyahu
charges
Anton La Guardia
Jerusalem
Israel's Prime Minister
Benjamin Netanyahu
his coalition => Nethanyahu's coalition
last night => night of 20-apr-1997
yesterday => 20-apr-1997
fraud
breach of trust

Using the conceptual network of the LMS (based on WordNet), terms can be expanded. For example, "Prime Minister" gives the following expansion:

SYNSET: Prime Minister, MP, Premier
HYPERNYM: head of state, chief of state
MEMBER OF: Cabinet

And "fraud" makes a link to "crime":

SYNSET: fraud
HYPERNYM: crime, law-breaking

On the basis of the expanded terms, a DIR is constructed that contains the most significant terms, but also an indication of the relevant domains. This is done by counting the score of the various terms to a certain domain. In the simple manual experiment we did with the text item above, the following domain-scores were obtained:

9 politics
6 Israel
4 justice

In this way, the DIR will easily match with a user profile expressing an interest in Israeli politics. The multi-lingual set-up offers some interesting possibilities. For example, a user from Spain subscribing to an English news wire service, may express his profile in Spanish, and browse through the Spanish concept space. Since the result is a set of concept identifiers, these can be matched with concepts derived from the English text. This will help Spanish users to read news items like the one expressed above using their own language.

We finally remark that WordNet is a useful resource, but it also has deficiencies. We have found that it runs often into problems of over-differentiation, i.e., too many word senses are distinguished. It contains numerous errors and omissions (for example, Britain is classified as a kingdom, but the Netherlands not). Similar findings have been done by other researchers. For that reason, we are currently working on a completely new concept structure, based on the WordNet resources and inheriting much of its structure, but set up in a more disciplined way.

8 SUMMARY AND FUTURE WORK

This paper described the fundamental aspects of a scalable, semantically oriented, 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 item and about its placement within a given document context. To surmount semantic-drifts and the terminology problem an 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.

Future work addresses the semi-automatic generation of link weights based on term co-occurrences and using statistical/probabilistic algorithms. In IR these algorithms use word and/or phrase frequency to match queries with terms. In addition we plan to concentrate on extracting more semantic knowledge form the link structure of documents. For example, in addition to terms from index records, the hyper-links (in case of HTML documents) may also provide useful information for the clustering process. Non-local hyper-links, i.e., hyper-links that are not parts of a single integrated document, are used to link documents that are content related. These can also be considered in the document clustering process when taking into account HTML based DL documents. And, finally, we are working at an improved and multi-lingual version of WordNet.

REFERENCES

C. M. Bowman et al. (1995), Harvest: A Scalable, Customizable Discovery and Access System, Univ. of Colorado - Boulder, CS Dept., techn. report CU-CS 732-94, (revised March 1995).

J. Carbonell et al. (1995) Translingual Information Retrieval: A comparative evaluation, *IJCAI '97*, Nagoya, Japan.

H. Chen (1994) Collaborative Systems: Solving the Vocabulary Problem, *IEEE Computer*, May.

Cycorp, Inc. Cyc Ontology (1995) *http://www.cyc.com/cyc-2-1/intro-pubic.html.*

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

N.V. Findler (1979) A Heuristic Information Retrieval System Based on Associative Networks, in *Associative Networks*, (ed. N.V. Findler), Academic Press.

P. Francis, T.Kambayashi, S. Sato, S. Shimuzu (1995) INGRID: A Self-Configuring Information Navigation Infrastructure, *4th Int'l WWW Conference Proceedings*, Boston, Ma.

J. Gilarranz, J. Gonzalo, F. Verdejo (1997) An Approach to Conceptual Text Retrieval Using the EuroWordNet Multi-Lingual Semantic Database, *Working Notes of AAAI Spring Symposium on Cross-Language and Text Retrieval*, Stanford Ca.

M. Hearst, J. Pedersen (1996) Reexamining the Cluster Hypothesis: Scatter/Gather on Retrieval Results, *ACM SIGIR'96 Conf.* , Zurich, Switzerland.

Y. W. Kim, J.H. Kim (1990) A Model of Knowledge Based Information Retrieval with Hierarchical Concept Graph", *Journal of Documentation*, vol.46, no.2.

M. Koster (1994) ALIWEB - Archie-like Indexing in the Web, *Procs 1st International Conf. on the World-Wide Web*, Geneva, Switzerland.

C. Lagoze (1996) The Warwick Framework: A Container Architecture for Diverse Sets of Meta-data, *Digital Libraries Magazine*, July/August.

R. Larson (1992) Experiments in Automatic Library of Congress Classification, *Journal of American Society for Information Science*, vol. 43, no. 2.

The Library of Congress (1996) Machine-Readable Cataloging, *http://lcweb.loc.gov/marc/marc.html.*

M.L. Mauldin, J.R. Levitt (1994) Web-agent related Research at the CMT, *Procs. ACM Special Interest Group on Networked Information Discovery and Retrieval (SIGIR'94).*

P. Mokapetris (1995) Domain Names - Implementation and Specification, *RFC 1035, anonymous ftp: ds.internick.net/rfc/rfc1035.txt.*

G. Miller (1995) WordNet: A Lexical Database for English, *Communications*

of ACM, vol. 38, no. 11.

S. Milliner, A. Bouguettaya, and M. Papazoglou (1995) A Scalable Architecture for Autonomous Heterogeneous Database Interactions, *21 Int'l Conference on Very Large Databases*, Zurich, Switzerland.

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

Y. Papakonstantinou, H. Garcia-Molina, J. Ullman — (1996) MedMaker: A Mediation System Based on Declarative Specifications *12th Int'l Conf. on Data Engineering*, New Orleanes.

M. Papazoglou (1995) Unraveling the Semantics of Conceptual Schemas, *Communications of ACM*, vol. 38, no. 9.

M.P. Papazoglou, S. Milliner (1996) Pro-active Information Elicitation in Wide-Area Information Networks, *Int'l Symposium on Cooperative Database Systems for Advanced Applications*, Kyoto, Japan.

B. Pinkerton (1994) Finding what People Want: Experiences with the WebCrawler, *Procs. 1st Int'l Conference on the WWW*, Geneva.

R. Rada et al. (1989) Development and Application of a Metric on Semantic Nets, *IEEE Transactions on Systems, Man and Cybernetics*, vol. 19, no. 1.

G. Salton (1989) Automatic Text Processing, *Addison-Wesley*, Reading Mass.

R.B Schatz et al. (1996) Interactive Term Suggestion for Users of Digital Libraries, *1st ACM International Conf. on Digital Libraries*, Bethesda MD.

R.B Schatz et. al. (1996) Federating Repositories of Scientific Literature: the Illinois Digital Library Project, *IEEE Computer*, May.

T. Smith (1996) The Meta-Data Information Environment of Digital Libraries, *Digital Libraries Magazine*, July/August.

S. Weibel, J. Goldby, E. Miller (1996) OCLC/NCSA Meta-data Workshop Report, *http://www.oclc.org:5046/oclc/research/conferences/ metadata/dublin_core_report.html.*

H. Weigand (1997) A Multi-lingual Ontology-based Lexicon for News Filtering - The TREVI project *IJCAI-Workshop Ontologies and Multilingual NLP*, Nagoya, Japan.

R. Wiess, et al. (1996) HyPersuit: A Hierarchical Network search Engine that Exploits Content-link Hypertxet Clustering, *7th ACM Conf. on Hypertext*, Washington DC.

P. Willett (1988) Recent Trends in Hierarchical Document Clustering, *Information Processing and Management*, vol. 24, no. 5.

9 BIOGRAPHY

Michael P. Papazoglou is a full Professor and director of the INFOLAB at the Univ. of Tilburg in the Netherlands. His scientific interests include cooperative information systems, heterogeneous database systems, object-oriented systems, distributed computing, digital libraries, electronic marketing and commerce. Papazoglou is the founding editor and co-editor in charge of the International Journal of Cooperative Information Systems and serves on several committes and advisory boards of several international journals. He has served as general and program chair for a number of well-known international conferences. Papazoglou has authored or edited eight books and approximately 100 journal articles and numerous refereed conference papers and given invited lectures in several countries.

Hans Weigand studied Computer Science at the Vrije Universiteit in Amsterdam, the Netherlands. In 1989, he moved to Tilburg University and worked on an ESPRIT project in document databases for two years. Since then he is Assistent Professor at the Department of Information Systems and teaches courses in the area of Computer Networks, Programming and Group Support Systems. His research is focused on the use of linguistic instruments in knowledge engineering; the development of a formal language for business communication based on speeach act theory and logic; and in natural language descriptions for object-oriented methods. He has been involved as a key person in several ESPRIT projects at the INFOLAB at the University of Tilburg. Weigand has published more than 30 refereed papers in international journals and conference proceedings.

Steven Milliner is currently a PhD student at the School of Information Systems at the Queensland Univ. of Technology (QUT) in Brisbane Australia. He has obtained his BSc degree in Computer Science from the University of Queensland and his Master's degree also in Computer Science from QUT. Milliner's interests are in interoperable databases, internet searching, networking and performance evaluation. He has published several papers in high profile conferences and has three papers in international journals. He is currently an Associate Lecturer at QUT and is expecting to complete his thesis in 1998.