Knowledge Navigation in Networked Digital Libraries*

Mike P. Papazoglou and Jeroen Hoppenbrouwers

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

Abstract. Formulating precise and effective queries in document retrieval systems requires the users to predict which terms appear in documents relevant to their information needs. It is important that users do not retrieve a plethora of irrelevant documents due to underspecified queries or queries containing ambiguous search terms. Due to these reasons, networked digital libraries with rapid growth in their volume of documents, document diversity, and terminological variations are becoming increasingly difficult to manage.
In this paper we consider the concept of knowledge navigation for federated digital libraries and explain how it can provide the kind of intermediary expert prompting required to enable purposeful searching and effective discovery of documents.
Keywords: digital library, meta-data, ontology, clustering, browsing, navigation, semantic indexing, concept searching.

1 Introduction

Digital libraries bring large volumes of information to the user, whether researcher, analyst, student or casual browser. The classical approach by Information Retrieval (IR) is to define scalable techniques such as the vector-space model for matching queries against many thousands of documents efficiently [21]. This technique attempts to maximize the relevance of a document to a query. AI approaches have also been similarly intensioned although focusing on applying domain knowledge and analogical reasoning rather than numeric matching techniques. For example, an analogical reasoning system can be used to construct the possible interpretations of query terms corresponding to alternative paths in the inference network and to negotiate them with the user. In this way the user is able to select his/her intended interpretation of an unstructured query.

The relevance of query terms to documents is only one part of a complex problem. Currently, there is a massive investment world-wide in making digital document repositories accessible over networks. The result of this is that users

* This research has been partially funded by the European Union under the Telematics project Decomate LIB-5672/B.

of Digital Libraries (DLs) are overwhelmed by the amount of documents that are required to assimilate but also of the constant influx of new information. At the same time there is also a major investment in providing indexing, categorization, and other forms of meta-data for DL documents and a large number of IR techniques have been developed for automatic categorization of repositories for which human indexing is unavailable. These activities result in quite diverse meta-data vocabularies, e.g., index and thesaurus terms, that characterize documents. Therefore, the number of meta-data vocabularies that are accessible but unfamiliar for any individual searcher is increasing steeply.

1.1 Limitations of Index Terms

Despite user knowledge that several terms within a particular domain may have the same meaning, known IR technology can only match terms provided by the searcher to terms literally occurring in documents or indexing records in the collection. Unfortunately, keyword expansion techniques have shown no significant improvements over other standard IR techniques as it is usually very difficult to choose which keywords to expand [5]. This implies that there are too many potentially matching documents which may not be retrieved due to the variation of the index terms used, and the fluidity of concepts and vocabularies in different domains.

The situation described above is particularly acute in digital libraries with spatial distribution which aim to make widely distributed collections of heterogeneous documents appear to be a single (virtually) integrated collection. Such federated digital libraries (FDLs) typically specialize in a fairly *narrow* and *specific domain area*, e.g., Biomedicine, Computer Science, or Economics. Although the amount of searching in FDLs is expected to rise, diminishing search effectiveness and less reliable answers is the predictable result as a consequence of the explosive increase in meta-data heterogeneity due to terminology fluctuations. The challenge is to provide automatically the kind of expert assistance that a human search intermediary, familiar with the source being searched, would provide. In [3] has been argued that the most effective solution to improving effectiveness in the search of digital repositories would be technology to assist the information searcher in coping with unfamiliar meta-data vocabularies.

1.2 From Terms to Knowledge

A particularly promising methodology for addressing these objectives is *knowledge navigation*. This methodology relies on the use of computer assisted support for acquiring and relating digital information originating from diverse heterogeneous document repositories. Knowledge navigation combines techniques from knowledge representation and natural language processing with classical techniques for indexing words and phrases in text to enable a retrieval system to make connections between the terminology of a user request and related terminology in the information provided in an FDL.

At this juncture it is useful to discriminate between *terms* and *concepts*. Terms may appear in documents or meta-data descriptions and may originate from a controlled vocabulary of terms such as a thesaurus and have a predominantly structural flavor. Concepts, on the other hand, are used to organize index terms into distinct, higher-level, conceptual categories that have a distinct meaning.

The purpose of knowledge navigation is to help users negotiate a pathway through an overwhelming universe of information in order to improve their understanding. This requires locating, identifying, culling, and synthesizing information into knowledge. Knowledge navigation does this by analyzing the conceptual structure of terms extracted from document indices and using semantic relationships between terms and concepts to establish connections between the terminology used in a user's request and other related terminology that may provide the information required.

We consider the development of a methodical, scalable search process critical to the successful delivery of information from networked digital library systems. Hence, in order to provide users with tools for knowledge navigation, a four step process may be introduced: (i) *Determining* the information needs of users by means of different term suggestions; (ii) *Locating* candidate documents that may address these needs; (iii) *Analyzing* the structure, terminology and patterns of use of terms and concepts available within these information sources; and finally, (iv) *Retrieving* the desired documents. The very nature of this process suggests that we should provide facilities to landscape the information available in FDLs 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.

To support the process of knowledge navigation while overcoming the complexity of wide-area information delivery and management, we cannot rely on a collection of meta-data index terms which simply contain terms reflecting the content of documents in an FDL. A more structured and *pro-active* approach to searching is required. In such situations, *concept browsing* can be particularly beneficial [10, 18]. The precursor of such an advanced browsing approach assumes that we are in a position to impose some logical organization of the distributed information space in such a way that potential semantic relationships between related documents in the network can be explored. Accordingly, the objective of knowledge navigation systems is to be able to handle a spontaneous description of the information required while minimizing the need for an information seeker to engage in repeated query reformulation in order to discover the exact terminology that will retrieve the information required.

In this paper we discuss how the use of knowledge knowledge navigation techniques can be used to transform an FDL from a passive warehouse of navigatable information to an environment that supports pro-active distributed document searching and retrieval. The paper is organized as follows. First we introduce the precursors of knowledge navigation such as a common ontology and rich meta-data sets. Following this, we discuss the benefits of knowledge navigation for FDLs and introduce a conceptual FDL architecture. Subsequently, we discuss

different dimensions of browsing and querying and report on related research. Finally, we summarize the main points of this paper.

2 Conceptual Network Creation and Maintenance

For knowledge navigation to be effective it should provide an efficient network of pathways that can allow a person to navigate through conceptual space in a DL and can also reveal relationships between concepts. It should support human browsing and navigation in "conceptual space" by providing a structured map of the concepts used in the indexed material and allowing a user to move conveniently back and forth between concepts in a classification scheme and thus locate the text material where these concepts occur. It should also be able to use paths in the conceptual index to find relationships between terms in a request and related terms that may occur in relevant material. In the following we present some relevant terminology and explain why the use of ontologies and conceptual networks can be beneficial to knowledge navigation.

2.1 From Indexing Terms to Conceptual Networks

Indexing terms are used when adding a document to a (digital) library for efficient retrieval of the document. Surrogates of the documents in a digital library, commonly known as meta-data, are created by professional catalogers and indexers. The concept of meta-data is examined further in section 3.

Vocabulary in information retrieval usually refers to the stylized adaptation of natural language to form indexing terms. In such situations we tend to define a vocabulary purely in terms of word structures that can be manipulated, but the meanings of the words are constructed subjectively and situationally and the use of the vocabulary is predominantly social [3].

A thesaurus in the field of information and library science is defined as "a compilation of words and phrases showing synonyms, hierarchical and other relationships and dependencies, the function of which is to provide a standardized vocabulary for information storage and retrieval systems" [20]. Such a list of thesaurus terms, also called an authority list, is useful in showing terms, which may be used in indexing, and which should be not.

Conventional thesauri often represent a general subject area, so that they usually need significant enhancement to be tailored to a specific domain. This has triggered AI research to attempt to represent knowledge of a domain in a declarative formalism, with the goal of permitting knowledge to be expressed with such detail that it can be manipulated automatically.

An ontology may be generally defined as a representation of a conceptualization of some domain of knowledge [8]. It is a formal and declarative representation which includes the vocabulary (or names) for referring to the terms in that subject area and the logical statements that describe what the terms are, how they are related to each other, and how they can or cannot be related to each other. Ontologies therefore provide a formal vocabulary

for representing and communicating knowledge about some topic and a set of relationships that hold among the terms in that vocabulary. This consensus knowledge about a specific and narrow domain is meant to be relatively stable over time, and reusable to solve multiple problems.
Formal ontologies define vocabulary with logic. The exact syntax and semantics depends on the representation language, e.g., description logics [27]. Formal ontology concept definitions are usually constructed as frames with definitions including a name, a set of relations to other concepts, and a natural language description that serves strictly as documentation [27].
Informal ontologies, such as WordNet [13], use a dictionary style natural language description, and this description provides the authoritative meaning of the term. Informal ontologies use richer kinds of relationships than subsumption and are directed graphs rather than trees as in the case of formal ontologies.
Compared to description systems in DLs, ontologies are more expressive, precise and powerful. They are powerful because their precision supports reasoning. Ontologies can be used to define sets of of descriptive meta-data, e.g., the Dublin Core elements, see section 3, as well as systems for classifying knowledge [27].

A conceptual network is a collection of semantic nodes with links between them, in such a way that many relationships are captured. Detailed coverage of a domain is an elaborate process involving rich semantic relationships, e.g., semantic roles and part-of relationships [13], usually more than those that a typical thesaurus can sustain. However, newer generation thesauri and ontologies contain richer information that can be used as basis to construct conceptual networks [14].

As the vocabulary of each living language grows continuously, especially in the technical-scientific domains, it will be very hard to claim that *any* thesaurus is ever complete. Regular updates must be applied to every thesaurus to keep it abreast of terminology evolution and changes [1].

2.2 Managing Network Growth

The dynamic nature of thesauri and conceptual networks means that most static, hierarchically organized classifications such as the UDC tree[1] or the classification of the Journal of Economic Literature (JEL)[2] are not adequate to serve as a complete conceptual network. More specifically, classification tools do not aim at covering the complete terminology of a domain, instead they aim to identify specific subfields (subjects) within broader fields. Of course their subject headings can be used as a starting point for thesaurus construction, and they can be included as generic 'see also' (related term) pointers in a conceptual network.

Conceptual networks such as WordNet [13] contain enough terminology and relationship information to be usable. However, these are usually too static and

[1] `http://main.bib.uia.ac.be/MAN/UDC/udce.html`
[2] `http://www.econlit.org/elclasbk.htm`

cover a broad range of common fields while being sparse on specialized domains – which are far better suited to assist users in knowledge navigation [2, 11]. It is especially important to have the conceptual network organized in terms of concepts instead of plain index terms. WordNet uses the *synset* primitive to group highly synonymous terms together while the EuroWordNet project extends the synonymy relation to include multiple languages [24, 25]. Other work on Lexicons, aimed specifically at conceptual modeling [9], also suggests ways of organizing terminology to properly present a conceptual space to users.

Acquiring a suitable conceptual network therefore is not just a matter of copying existing thesauri or term lists. Considerable effort should be put into the creation and maintenance of a conceptual network for knowledge navigation purposes. Any semantic network which models a piece of reality needs regular updating in order to stay synchronized with the world it represents. It is unreasonable to expect that a network can be constructed once and remain stable for an extended period of time. According to [16]: "The danger is that if the thesaurus is permitted to become monolithic and resistant to change, it can actually hinder both indexing and retrieval."

In the case of a virtual library system – which exhibits spatial distribution and which specializes in one particular scientific field, such as economics, astronomy or chemistry – the network should be maintained by experienced librarians and catalogers. These people can quickly recognize the particular places in the conceptual network where potential new concepts should be placed, and can update and verify the network as part of their regular work. In this way they help develop a 'conceptual map' of their domain, which can be very useful for other purposes besides knowledge navigation support.

3 Meta-data: the Foundations of Document Description and Discovery

Surrogates of the documents in a digital library – called *document index records* (DIRs), or *meta-data* – are usually created by professional catalogers and indexers. 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 [22]. The purpose of meta-data is to describe a certain the type of a resource and provide the means of identifying topics related to the search terms.

In recent years there has been a focus on meta-data in relation to describing and accessing information resources through digital libraries, or the World Wide Web in general.[3] In contrast to traditional descriptive cataloging, which relies on very complex rules requiring extensively trained catalogers for successful application, simpler descriptive rules are employed which are sufficiently simple to be understood and used by the wide range of authors and publishers who

[3] `http://ifla.inist.fr/II/metadata.htm`

contribute information to the Web. Many librarians and organizations create handicraft collections of records (portals) that are more informative than an index entry but is less complete than a formal cataloging record to characterize document resources. Some of these collections of "third-party" meta-data records classify the document resources using organizational methods such as the Library of Congress classifications, UDC codes, or home grown schemes. The collections also include subject or keyword information, as well as title and authority information.

The term meta-data in the context of DLs has been used in conjunction with the "Dublin Core" [26] which is being developed as a generic meta-data standard for use by libraries, archives, government and other publishers of online information. The Dublin Core was intended to be limited to describing "document like objects" such as HTML pages, PDF files and graphic images. It was intended to be descriptive, rather than evaluative. The Dublin Core standard was deliberately limited to a small set of elements which would have applicability over a wide range of types of information resources. However, the descriptive rules suggested by the Core do not offer the retrieval precision, classification and organization that characterizes library cataloging.

To support pro-active searching FDLs need to rely on higher-level (and more structured) meta-data than that of descriptive cataloging to support dealing with the problems of large-scale searches and cross disciplinary semantic drifts. The meta-data *schema*[4] should capture in its fields the contents and topics of documents based on elements of the Dublin Core, e.g., title, creator, subject and textual summaries (description), and also provide fields that allow to associate search terms and concepts to related sets of terms and topics in other documents. It is particularly useful to be able to combine meta-data descriptions with ontologies. If an ontology underlies meta-data descriptions, then it can represent the meta-data terms associated with documents in a precise and explicit manner. It can help alleviate term mismatch problems by grounding meta-data supplied terms to commonly used and understood terms. It can also ontologically define implicit (narrower, broader, part of) relationships between meta-data supplied terms, thus, making them amenable to computational reasoning.

4 Requirements for Effective Knowledge Navigation

It is evident that facilitating access to a large number of distributed document repositories and libraries involves a range of requirements that cut across both user and system needs.

Topic classification schemes In order to be able to search large information spaces an important requirement is to partition them into distinct subject (topic) categories meaningful to users. This makes searches more directed and efficient. It also facilitates the distribution and balancing of resources via appropriate allocation to the various partitions.

[4] `http://www.imsproject.org/md_overview.html`

Abstracting meta-information Support for meta-information concentrates not on the descriptions (meta-data) of network-accessible information items but rather on high-level information whose purpose is to cross-correlate, collate, and summarize the meta-data descriptions themselves. This type of summarization or synoptic topic knowledge is called *meta-information*. Thesaurus-assisted explanations created for each such subject-based abstraction (and its contents) can serve as a means of disambiguating term meanings and addressing terminology and semantic problems.

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 a knowledge navigation 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.

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 meaningful 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, viz. index terms.

Relevance feedback and results explanation The need to provide information users with explanations regarding the rationale for the relevance of information presented in response to queries and of the meanings of the terms occurring in the presented information is apparent.

Scalability support Scalability is an important issue for any large distributed system as it deals with the management of distributed resources, repositories, and document collections. A scalable system is one that can grow piecemeal without hindering functionality or performance if the current system configuration expands beyond the resources available.

5 Federating Digital Libraries

The issues presented in the previous sections illustrate the wide range of problems to be considered when designing and implementing an FDL. This section presents a conceptual architecture for an FDL and illustrates how issues identified in the previous can have implications in several areas of this architecture.

In the following we will describe two different approaches to the problem of federating digital libraries. The first approach is based on the premise that the interconnected DLs agree on using a single standard ontology (or thesaurus) for cooperation. The second approach is based on the premise that although

individual DLs agree on cooperating they wish to retain complete control and autonomy of their local thesauri – which can also continue to evolve with the passage of time. This second configuration is typical of cases where there is an element of multi-linguality involved.

In both cases our approach to knowledge navigation in FDLs is based on linguistic techniques and ontology-based categorization. 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. Prior to describing the two different configurations to federating DLs we will describe a conceptual architecture for FDLs which will be used as a reference to explicate their differences.

5.1 Conceptual Architecture for Federated Digital Libraries

To exemplify the FDL environment we use a comprehensive example from a federated library in Economics embracing various institutional libraries scattered over the European continent. Each library maintains its own collection of documents, using both full text and controlled vocabulary indexing. Users of the FDL in Economics should be able to search and access documents no matter where they originate from and irrespectively of the terms used to index the documents in the individual libraries.

Figure 1 shows a conceptual view of this FDL. The architecture is in a position to provide a conceptually holistic view and cross-correlate information from the multiple libraries (repositories). The in the FDL meta-data schemas contain meta-data terms in addition to other descriptive information such as geographical location of documents, access authorization and usage roles, charge costs, and so on. An aggregation of meta-schema terms for semantically related documents will result in forming a subtopic. For example, meta-data schemas individual libraries may abstract documents about market structure and pricing and may contain such index terms as monopoly, oligopoly, auction, rationing, licensing, etc. The aggregation of these terms generates a more generic subtopic (*concept*) called Market Models, step 2 in Figure 1. Although this concept is semantically clear to many users, it is highly unlikely that the term 'market models' appears as such in the documents. Finally, semantically related concepts such as Industrial Economics, Household Economics, Consumer Economics and Market Models are aggregated in their turn into the higher-level concept Micro-Economics, see step 3 in Figure 1. We refer to this type of construct as *Topic* or *Generic Concept* [19]. In this example, we assume for reasons of simplicity that terms are connected to topics via a single level of concepts. However, in a reality terms may be connected to topics via an elaborate hierarchy of concepts.

Topics 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 topic is thus a form of a logical object (a kind of *a contextualized*

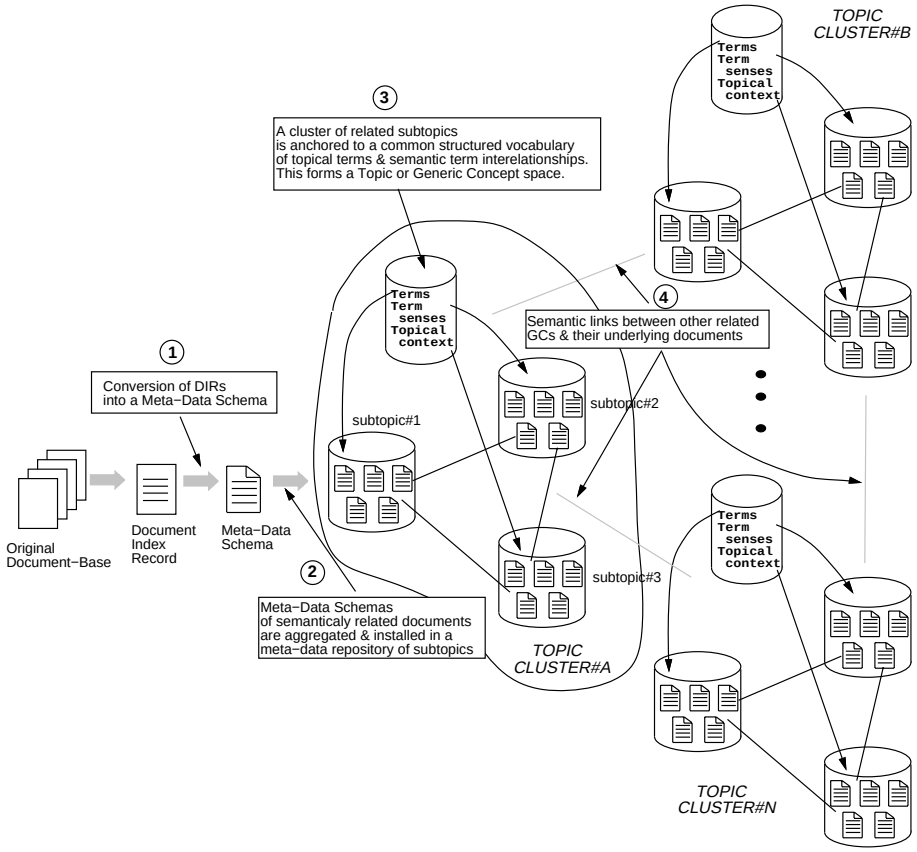

Fig. 1. Connecting meta-data schemas and forming the Topic space.

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.

Overall a networked digital library system (representing a narrow domain, e.g., economics, astronomy or engineering) may be viewed in terms of four logical layers, as depicted in Figures 1 and 2, where

1. the top most layer corresponds to the *topic or generic concept* layer;
2. the second layer from the top represents the *subtopic or concept* layer associated with the meta-data schemas;
3. the third layer represents the *index terms* associated with the documents;
4. the bottom layer corresponds to the *document collection* layer (document base in Figure 1).

This four-tier architecture is the key ingredient to knowledge navigation 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 subtopic layer, and finally to the generic 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. the topic layer, and percolate to the schema layer in order to provide access to the contents of a document cluster. This methodology results in a simplification of the way that information pertaining to a large number of interrelated collections of documents can be viewed and more importantly it achieves a form of global visibility.

This type of topic-based clustering of the searchable information space provides convenient abstraction demarcators for both the users and the system to make their searches more targeted, scalable and effective. This type of subject partitioning creates smaller semantically related collections of documents that are more efficient for browsing and searching. Concept searching can be utilized as opposed to keyword searching which is the traditional method employed by most contemporary search engines.

5.2 Tight Coupling: a Common Ontology-based Approach

The tightly coupled architecture describe in this section is based on earlier research activities on the TOPICA federated digital library system [19]. The architecture has as its main objective to impose a logical order to an otherwise flat information space by categorizing the content of document meta-data schemas and clustering them into topically-coherent, disjoint groups which are anchored on standard ontologies. Classical document clustering techniques from IR are used for this purpose [15]. The information space in FDLs is logically partitioned into meaningful subject areas. This results in clusters of documents formed around specific topic categories where different kinds of term suggestions – automatically generated by a thesaurus (ontology) – can be used to enhance retrieval effectiveness. We refer to this setup as the *topic space* for each group of semantically related documents, see Figure 2. After individual contextual spaces of documents are formed, subject-specific browsing or searching can be performed by a variety of tools that concentrate on concept (as opposed to term) browsing. Only in this way we can allow tools and searchers to selectively access individual document aggregations while ignoring others. The inclusion of a complete vocabulary and semantic information in the topic space provides the opportunity for "intelligent" navigation support and retrieval, with the system taking a more active role in the navigation process rather than relying purely on manual browsing.

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 topic. A common ontology is used to disambiguate topic-related terms and concepts and terms originating from different meta-data sets in the networked DLs. The common (canonical) ontology, e.g., an appropriate extension of the in-house Attent thesaurus, is used to represent concepts, terms and their relationships in a conceptual graph structure, akin to an

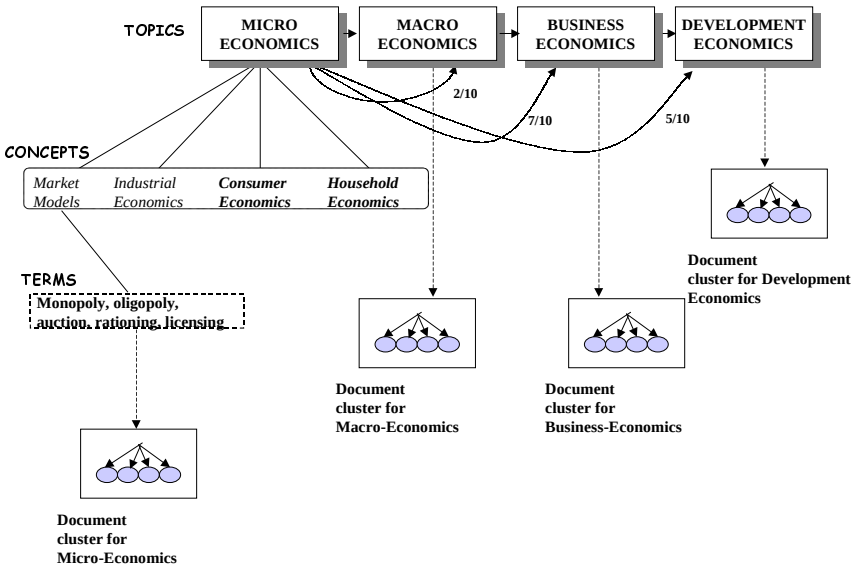

Fig. 2. Forming a conceptual network by linking documents to concepts and topics.

associative thesaurus. Term disambiguation for the diverse meta-data terms and their surrounding concepts is achieved with reference to this conceptual network to make connections between requested items and indexed terms of information.

A topic is materialized by a class hierarchy depicting all concepts and terms sampled by the topic, e.g., Micro-Economics. Each topic is characterized by its name and the context of its concepts and terms. A topic's concept space consists of abstract descriptions of terms in the domain, ontological relationships between these terms, composition of terms, terminology descriptions, hypernym, hyponym, antonyms-of, part-of, member-of (and the inverses), pertains-to relations, selected term usage and definitions (narrative descriptions), domains of applicability, list of keywords, and other domain specific information that apply to the entire collection of members of a topic. For example, if the user chooses to explore the topic Micro-Economics (s)he will view the terms shown by the concept browser on Figure 3. Once the concept Household Economics has been selected then a term bucket containing all possible terms under this topic is revealed. Subsequently, the user is free to choose terms that reflect her/his own preferences to form queries against the entire FDL. Terms in documents are matched to those appearing in the term bucket by word analysis techniques [15]. Hence, the user is pointed to the relevant documents where in the first instance (s)he can see (and possibly query) the document meta-data schema. The topic-areas, described by the topic descriptor classes, are interconnected by weighted links to make the searches more directed, see Figure 2. When dealing with a specific

concept such as Market Models we are not only able to source appropriate information from remote document-based on the same topic but also to provide information about semantically related topics, e.g., Business Economics in the case of the Micro-Economics topic. The stronger the weight the closer the relatedness between two topics. Documents within a topic are all connected to this topic by a weight 10/10. Currently, the weights to topics are manually assigned by catalogers. This can be replaced in the future by text analysis techniques and IR ranking algorithms to determine the relatedness of topics.

In summary, the topic structure is akin to an associative ontology (thesaurus) and on-line lexicon (created automatically for each topic category). Ontology-assisted explanations created for each topic-based information space serve as a means of disambiguating term meanings, and addressing terminology and semantic problems. Therefore, the topic structure 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.

5.3 Loose Coupling: Inter-linking Independent Thesauri

One problem with the approach outlined above is that an agreed upon conceptual network needs to be maintained on the basis of a common ontology (thesaurus). In many cases, the individual libraries contributing to the virtual library will demand complete freedom in maintaining their own, specialized, localized system, including the index vocabulary (thesaurus). However, these libraries would not object against re-using their thesauri, and would favor mutual linking of concepts between thesauri. In such cases we need to provide software solutions that permit users to pose queries using terms from a thesaurus (source thesaurus) that was not used to index the documents being searched. A *cross-thesaurus gateway* will then translate the query into terms from the remote thesaurus (target thesaurus) that was used to index the documents. We will explain this approach based on our experience with working on the European virtual library for Economics.

The Decomate Project[5] is an example of a truly federated, virtual library for Economics. The contributor libraries are geographically distributed over Europe and each partner maintains several databases, indexed using different thesauri, e.g., EconLit/JEL, IBSS, Attent, in different languages (English, Spanish, and Italian).

In Decomate, a Multi-Protocol Server is capable of simultaneously querying all relevant thesauri: a 'horizontal' multi-query can be issued that retrieves all matching terms out of all thesauri. Decomate does not directly support integration of the federated thesauri. However, it provides a cross-thesaurus linkage (bridging) facility which allows generating a virtual concept network involving terms from any two interacting thesauri based on semantic closeness, see Figure 4. A connection can still be made if we follow neighboring, viz. semantically related, concepts in the conceptual network which may lead to matching concepts in the thesauri. When concepts are semantically matched, the terms contributed

[5] `http://www.bib.uab.es/decomate2`

by all thesauri can be collected in a virtual term bucket, originating from the meta-data underlying the matched documents, in order to facilitate the accessing of documents whose terms are missed by the indexers (Figure 4).

Some thesauri (such as JEL) include unique codes for concepts. For example, Household Behavior: General has the JEL code D10, irrespective of the actual term or language used for its description. Related JEL codes are D11 Consumer Economics: Theory, D12 Consumer Economics: Empirical Analysis, and D13 Household Production. Linking up such instances of the JEL thesaurus in different languages is therefore an easy task.

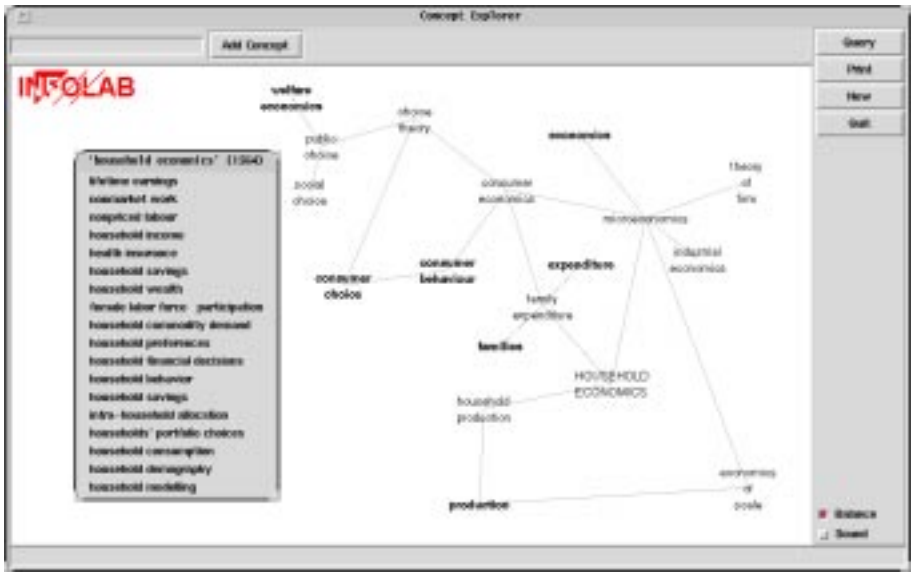

Fig. 3. Browsing the Attent Thesaurus

The virtual conceptual network is, just like a view in the database parlance, created dynamically, and in bottom up fashion, every time a user fires a query containing a term that matches a local thesaurus. This is contrast to the approach taken in section 5.2 where a a fixed ontology is used as a basis for matching concepts from different DLs in a top down fashion. The virtual conceptual network is not only used for concept matching but also for user browsing purposes.

6 Information Discovery Strategies

An interesting dichotomy in the space of document retrieval strategies is the distinction between *searching* and *browsing*.

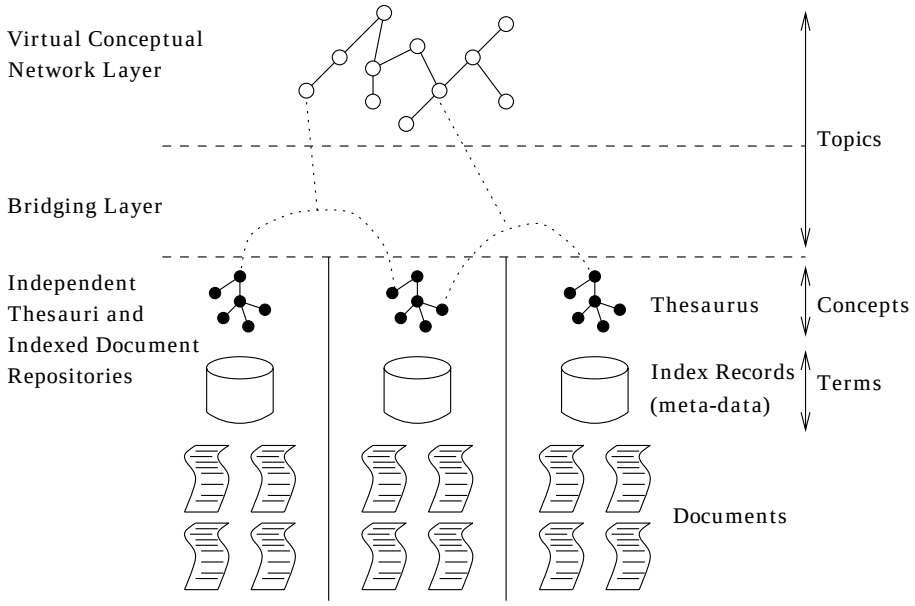

Fig. 4. Using a bridging layer to create a virtual conceptual network of concepts and related terms

Searching implies that the searcher knows exactly what s/he is looking for. If the collection to be searched is small *compared to the precision of the query*, the resulting number of 'hits' will be sufficiently small to allow further processing. Searching falls short, however, when the user is required to know (or remember) the valid keywords, how these keywords correlate with concepts that s/he wishes to find, and how the keywords may be combined to formulate queries.

Traditional IR queries are considered as an analytical strategy, requiring planning, cognitive overhead, goal-driven and batch-oriented techniques. However, when faced with ill-defined problems requiring information access, users often wish to explore the resources available to them before exploiting them. This exploration may be partly aimed at refining their understanding of the potential information space or content that is available to, and partly aimed at formulating a concrete course of action for retrieving specific documents. Tools that support the browsing of document meta-data collections, as opposed to searching, are aimed at satisfying this need to learn more about documents in a collection before taking any action.

The purpose of browsing is to provide an open, exploratory information space to the user. Browsing can be accomplished by providing links between terms that can be explored at will as the focus of exploration changes. In many cases as new information is obtained in the process of browsing the goal may change. Strategies can be selected in response to these conditions to pick up new chunks of information. We can view browsing as a semi-structured, heuristic, interac-

tive and data-driven activity of exploratory nature quite distinct from keyword (boolean) searching. Navigation can be seen as a special form of browsing characterized by high interactivity in a structured environment with the destination seldom predetermined. Navigation balances user and system responsibility with the user making choices from directions provided by the system. Navigation provides "pathways to discovery instead of answers to queries" [6]. Therefore, navigation is an ideal guide for *serendipity of information*, where users browse at random seeking information that is unknown, often not knowing what their target is unless it is seen.

Knowledge navigation is an advanced form of navigation where the system plays a more pro-active role by locating, identifying, culling, and synthesizing information into knowledge that it uses to assist the information seeker to discover the exact terminology that will retrieve the information required. It is not surprising that knowledge navigation concentrates on browsing prior to embarking on searching (querying) activities. In this way searches become more directed and effective as unwanted material is discarded during the process of navigation.

In section 5.2 we explained how navigation can be used to guide the user to discover the exact terminology required to retrieve documents dealing with specific issues under the broader topic of Micro-Economics, see also Figures 2 and 3. This is one form of navigation that can be provided with the FDL configurations described in section 5. We refer to this mode of navigation as *index-induced* navigation.

Another form of navigation that can be used with systems that provide weighted relationships among topics, see Figure 2, is that of *topic-driven navigation* which is when the user embarks on explorative searches and is most likely interested to find data closely related to a local document by following topic link-weights. We will use the topic connections shown in Figure 2 to illustrate this form of navigation. The concept-driven search is based on the weights with which a specific document base, e.g., Market Models – which is the subject of interest of some users – is linked to the various other topics in the system. This document base's weight to the Micro-Economics (its own topic) is 10/10, whereas its links to the topics Macro-Economics, Business Economics, and Development Economics are weighted with 2/10, 7/10 and 5/10, respectively. The Micro-Economics topic is in closer proximity to the Market Models document-base, followed by the Business Economics, Development Economics, and Macro-Economics topics. The user may then choose to explore concepts and meta-data information contained in the Micro-Economics topic first. Subsequently, s/he may choose to explore the Business Economics topic, followed by the Development Economics, and so on. The two modes of navigation can be mixed: when exploring these topics the user may embark on index-driven navigation to gain more insight into the concept found.

When the user needs to further explore the search target, *intensional*, or schema queries [17] – which explore meta–data terms – can be posed to further restrict the information space and clarify the meaning of the information items under exploration. Sample intensional queries related to the topics in the previous sections may include the following:

query-1: *Give me all terms similar to "value theory" under JEL AND Attent.*
query-2: *Give me all terms more specific than "value theory" and all their parts under JEL.*

The previous two queries return definitions and connections between concepts and terms under different thesauri.

Finally, when the users are sufficiently familiar with the terminology and understand the uses of the terms employed in an FDL they can issue *extensional queries* which retrieve documents or document meta-data (in case of non-electronic documents). Some representative extensional queries may be:

query-3: *Give me all documents dealing with "Household Behavior: General" under JEL AND "Family Expenditure" under Attent.*
query-4: *Give me all documents similar to author = "S. Hochguertel" AND "A. van Soest" AND title = "The relation between financial and housing wealth of Dutch households".*

Query-3 returns documents which belong to the intersection of two concepts in two different thesauri, wile query-4 tries to match a certain book pattern (through its associated meta-data) to that of other documents.

7 Related Work

Related work can be broken into two broad categories. First, work that spans different IR techniques such as query modifications and query refinement and clustering techniques. Second, activities in the area of digital libraries that concern themselves with subject-based information gateways.

7.1 Query Modification and Refinement

Related work on query modification has focused on automatic query expansion [7, 4] by means of addition of terms to a query to enhance recall. Query expansion has been done using thesauri or based on relevance feedback. Automatic query expansion techniques rely mainly on fully automatic expansion of terms to the query according to a thesaurus with no user intervention. The thesaurus itself can be either manually or automatically generated. With relevance feedback [4] query terms are selected or weighted based on a retrieved result set where terms are added to the query based on evidence of usefulness. Interactive query expansion can be used on basis of relevance feedback, nearest neighbors and terms variant of the original query terms that are suggested to the user.

Query refinement tries to improve precision (and not recall) by perusing the documents and selecting terms for query expansion which are then suggested to the user [23]. Automatically generated thesauri are used for suggesting broader and narrower search terms to the user.

Our approach differs from these activities as we place emphasis on characterizing document sets, logically partitioning them into distinct sets and then

interactively querying these sets based on concept rather than term retrieval. As basis of comparison we use a standard ontology. In this way users are assisted to formulate meaningful queries that return a large number of desirable documents.

7.2 Clustering Techniques

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 [28]. 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, follow 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.

7.3 Subject-based Information Gateways

Of particular interest to our work are *subject gateways*. These are facilities that allow easier access to network-based information resources in a defined subject area [12]. 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. Entries are usually created by a cataloger manually by identifying a resource, describing the resource in appropriate template which is submitted to the database for indexing.

Typical examples of subject gateways are: the Social Science Information Gateway (SOSIG),[6] which incorporates a complete thesaurus containing social science terminology, and the Organization of Medical Networked Information (OMNI)[7] 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

[6] `http://www.sosig.ac.uk/`
[7] `http://omni.ac.uk/`

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 to 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. In this way the resource containing numerous pages can be returned as an individual hit even by a search engine that indexes each Web page as a distinct entity.

8 Summary

In this paper we presented the concept of knowledge navigation for federated digital libraries and explained how it can provide the kind of intermediary expert prompting required to enable purposeful searching and effective discovery of documents.

We have argued that knowledge navigation in federated digital libraries should be 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 and enhance document retrieval, alternative search concepts and terms and terms senses are suggested to users. Finally, we have briefly outlined two FDL architectures, that are currently under development, which enable users to gather and rearrange information from multiple digital libraries in an intuitive manner.

References

1. J. Aitchison and A. Gilchrist. *Thesaurus Construction*. Aslib, London, 1987. 2nd edition.
2. R. Bodner and F. Song. Knowledge-based approaches to query expansion in information retrieval. In *Lecture Notes in Computer Science*, volume 1081, pages 146–158. 1996.
3. M. Buckland et al. Mapping entry vocabulary to unfamiliar meta-data vocabularies. *Digital Libraries Magazine*, Jan. 1999.
4. C. Buckley et al. Automatic query expansion using SMART. In *3rd Text Retrieval Conference: TREC-3*, Gaithersburg, MD, Nov. 1994.
5. J. W. Cooper and R. J. Byrd. Lexical Navigation: Visually Prompted Query Expansion and Refinement. In R. B. Allen and E. Rasmussen, editors, *Proceedings of the 2nd ACM International Conference on Digital Libraries*, 1997.
6. D. Cunliffe, C. Taylor, and D. Tudhope. Query-based Navigation in Semantically Indexed Hypermedia. In *ACM Hypertext Conference, Southhampton*, June 1997.
7. E. Efthimiadis. A user-centered evaluation of ranking algorithms for interactive query expansion. In *16th Annual Int'l ACM SIGIR Conference on Research and Development in Information Retrieval, Pittsburgh, PA*, June 1993.
8. T. Gruber. Toward principles for the design of ontologies used for knowledge sharing. Technical Report KSL-93-04, Knowledge Language Laboaratory, Stanford Univ., 1993.

9. J. Hoppenbrouwers. *Conceptual Modeling and the Lexicon*. PhD thesis, Tilburg University, 1997. `http://infolab.kub.nl/people/hoppie`.
10. J. Hoppenbrouwers. Browsing Information Spaces. In J. Prinsen, editor, *International Summer School on the Digital Library 1998*, Tilburg, The Netherlands, 1998. Ticer B.V. `http://infolab.kub.nl/people/hoppie`.
11. H. Howard. Measures that discriminate among online searchers with different training and experience. *Online Review*, 6:315–327, 1992.
12. J. Kirriemuir, D. Brickley, S. Welsh, J. Knight, and M. Hamilton. Cross-Searching Subject Gateways—the Query Routing and Forward Knowledge Approach. *D-Lib Magazine*, Jan. 1998.
13. G. Miller. Wordnet: A lexical database for english. *Communications of the ACM*, 38(11), 1995.
14. U. Miller. Thesaurus Construction: Problems and their Roots. *Information Processing and Management*, 33(4):481–493, 1997.
15. S. Milliner, M. Papazoglou, and H. Weigand. Linguistic tool based information elicitation in large heterogeneous database networks. In R. van de Riet, J. Burg, and A. van der Vos, editors, *Applications of Natural Language to Information Systems*, pages 237–246. IOS Press/Omsha, 1996.
16. J. Milstead. Methodologies for subject analysis in bibliographic databases. *Information Processing and Management*, 28:407–431, 1992.
17. M. Papazoglou. Unraveling the Semantics of Conceptual Schemas. *Communications of the ACM*, 38(9), Sept. 1995.
18. M. Papazoglou. Knowledge Navigation and Information Agents: Problems and Issues. 1997.
19. M. Papazoglou, H. Weigand, and S. Milliner. TopiCA: A Semantic Framework for Landscaping the Information Space in Federated Digital Libraries. In *DS-7: 7th Int'l Conf. on Data Semantics*, pages 301–328. Chapman & Hall, Leysin, Switzerland, Oct. 1997.
20. J. Rowley. A comparison between free language and controlled language language indexing and searching. *Information Services and Use*, 10:147–155, 1990.
21. G. Salton. *Automatic Text Processing*. Addison–Wesley, Reading Mass., 1989.
22. T. Smith. The Meta-Data Information Environment of Digital Libraries. *Digital Libraries Magazine*, July/August 1996.
23. B. Velez et al. Fast and effective query refinement. In *20th 16th Annual Int'l ACM SIGIR Conference on Research and Development in Information Retrieval, Philadelphia*, July 1997.
24. P. Vossen. EuroWordNet: a multilingual database for information retrieval. In *Proceedings of the DELOS workshop on Cross-language Information Retrieval, March 5-7, 1997, Zürich*, 1997.
25. P. Vossen, P. Diez-Orzas, and W. Peters. The Multilingual Design of the EuroWordNet Database. In *Proceedings of the IJCAI-97 workshop Multilingual Ontologies for NLP Applications, August 23, 1997, Nagoya*, 1997.
26. S. Weibel, J. Goldby, and E. Miller. OCLC/NCSA Meta-Data Workshop Report. `http://www.oclc.org:5046/oclc/research/conferences/metadata/dublin_core_report.html`, 1996.
27. P. Weinstein. Ontology-based meta-data. In *21th Annual Int'l ACM SIGIR Conference on Research and Development in Information Retrieval, Pittsburgh, PA*, 1988.
28. P. Willett. Recent trends in hierarchical document clustering: A critical review. *Information Processing and Management*, 24(5), 1988.