Landscaping the Information Space of Large Multi-Database Networks

M.P. Papazoglou
University of Tilbug
INFOLAB
P.O. Box 90153
5000 LE Tilburg
The Netherlands
MikeP@kub.nl

H.A. Proper
ID Research
Groningenweg 6
2803 PV Gouda
The Netherlands
E.Proper@acm.org

J. Yang
Internet Marketplace
CSIRO, Mathematical & Information Sciences
GPO Box 664
Canberra ACT 2601
Australia
Jian.Yang@cmis.csiro.au

Version of July 26, 1999 at 9:49

Abstract

The promises of network-accessible information are increasingly difficult to achieve. These difficulties are due to a variety of causes, such as, the rapid growth in the volume of network-available information and the increasing complexity, diversity and terminological fluctuations of the different information sources available.

This paper presents a conceptual architecture for the organisation *information space* across collections of component systems in multi-databases that provides serendipity, exploration and contextualisation support so that users can achieve logical connections between concepts they are familiar with and schema terms employed in multi-database systems. Large-scale searching for multi-database schema information is guided by a combination of lexical, structural and semantic aspects of schema terms in order to reveal more meaning both about the contents of a requested information term and about its placement within the distributed information space.

1 Introduction

The dramatic growth in global interconnectivity has placed vast amounts of data within easy reach. At the same time it has made on-demand access to widely-distributed information a natural expectation for users.

A complicating factor is the difficulty in providing coherent access and correlation of information that originates from diverse and widely-distributed sources. This is an involved process, not only because of the sheer volume of information available, but also because of heterogeneity in naming conventions, meanings and modes of data usage. Differences in descriptions, abstraction levels, and precise meanings of terms being used in disparate sources do not yield well at all to automation. These problems are compounded by differences in user perceptions and interpretations, and variations that may occur at autonomous sources over time. This renders users presented with the problem of gaining adequate knowledge of a potentially huge, complex and dynamic system, in order to access and combine information in a coherent and logical manner. Yet multi-database systems demand from users *prior detailed knowledge* of the definition and uses of their underlying data [KS97]. This expectation on a user's intellectual capacities is quite unreasonable in the context of a large distributed information space.

The focus in multi-database systems is on query processing techniques and not on how to discover where the actual schema elements in the component systems reside. No particular attention is paid to how schema items are structured, what they mean and how they are related to each other across component database schemas. The user's perception of the information content in networked databases is that of a vast space of information in a large flat, disorganized set of database servers. In contrast to this, our approach to searches for widely distributed information concentrates on providing a dynamic, incremental and scalable logical organization of component database sources, and search tools that are guided by this organization.

We view user interaction with a multi-database space as comprising two major phases, the:

schema information discovery phase where users systematically explore the multi-database space to locate potentially useful databases, and the

distributed query/transaction phase where the requested data sets are retrieved from the candidate databases.

We consider the development of a methodical, scalable search process critical to the successful delivery of information from networked database systems. Hence, in order to provide users with tools for the logical exploration of distributed information sources a four step process, termed *information elicitation* is introduced which targets the schema information discovery phase. This process encompasses the following steps:

1. *Determining* the information needs of users by means of different term suggestions;
2. *Locating* candidate database sources that address these needs;
3. *Selecting* schema items of interest from these sources; and finally,
4. *Understanding* the structure, terminology and patterns of use of these schema items which can subsequently be used for querying/transaction purposes.

The very nature of this process suggests that we should provide facilities to landscape the information available in large multi-database networks and enable 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 information elicitation while overcoming the complexity of wide-area information delivery and management, we cannot rely on a collection of indexes which simply contain schema information exported by individual database sources. A more structured and *pro-active* approach to searching is required. The precursor of such an advanced search approach assumes that we are in a position to impose some logical organization of the distributed information space in such a way that potential relationships between the component database systems in the network can be explored. In addition, to maintain scalability, this must be achieved through a decentralized mechanism which does not proceed via a one step resolution and merging of system information into a single static monolithic structure as advocated by many conventional practices for integrating multi-database systems. These and related issues are addressed in this article.

This paper presents the concept of information elicitation for large multi-database networks. The paper is organized as follows. Section 2 presents related work. In section 3 a logical organization for the semantic cross correlation of meta-data information from component databases in a multi-database system is defined formally. This logical organization of meta-data forms the core of our conceptual architecture for information space. Section 4 presents clustering techniques that allow the information space to be populated with available database nodes. Navigation and querying mechanisms to navigate and query the resulting information space are provided in section 5. Finally, section 6 presents some experimentation results while section 7 presents our conclusions and future work.

This work is an extension and elaboration of some early ideas outlined in [MBP95] and [PM98]. In [MBP95] we concentrated on the organization of physical data sharing in large database networks, and described how physical data sharing ties in with a pre-cursor of the conceptual organization of the information space presented in this paper. In [PM98] we described techniques and algorithms used for the conceptual clustering of databases. In this paper we concentrate on an extension and formalization of these ideas and on the logical grouping of databases, according to subject and a common terminology context. We present navigation and querying techniques for understanding the semantic context of networked databases.

2 Finding Information: An Overview

This section starts by providing a broad discussion of the problem of finding information in vast information spaces. This is complemented by a discussion of a number of techniques from different fields for locating information are discussed.

2.1 The vastness of information space

The elementary building blocks of information space are the information assets themselves. The term *information asset* can be defined as:

> any distinct information baring entity that is accessible on a networked environment and which may be combined with other such entities connected to same network.

A definition that truly supports the open character of the network. Examples of information assets included in this definition are:

- Web pages (including free text, sound, images, and video fragments).
- Free-text databases, such as newsgroups, mailing list archives, etc.
- Digital libraries.
- Traditional (relational, object-oriented) databases.

This definition of information asset should give an indication of the potential vastness of information space. Although in this paper we confine ourselves to information assets in conjunction with multi-database networks, the techniques outlined herein can also be applied to other examples of information assets.

2.2 Dealing with the vastness of information space

Different techniques have been, and still are being, developed to deal with the vastness of information space.

Web-based searching

The growth in 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. These search engines tend to return many potentially relevant information assets. Users are still required to manually wade through large result sets in search of truly relevant assets.

Most recent approaches to World Wide Web (WWW) querying [AMM97], [LRO96] concentrate only retrieval based on the contents of an information asset 'as-is'. They naively assume that the user (or search engine) is explicitly aware of the structure, semantics and vocabulary differences of the information assets that are available to them. However, due to the multiplicity, complexity, and terminology fluctuation of the information available, such an assumption is not practical. Practical studies have shown that there is a critical mismatch between a user's and the Web's vocabulary [Sch96]. Picking the right terms depends on how intimate searchers are with the vocabulary use in documents they wish to retrieve.

Centralized index search engines such as Lycos [ML94], Web Crawler [Pin94] are manual indexing schemes that rely on techniques 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): as the scope of the index 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 content-based search engines such as the Content Routing System [She95] and Harvest [Bow95]. These systems generate summarized descriptions (*content labels*) of the contents of information assets. The Content Routing System creates and maintains indexes of widely distributed sites. In this distributed information retrieval system a collection of documents is described by means of a content label which in turn can be treated as a document and can be included in another collection. Content labels help users explore large information spaces. However, document collections and their labels are confined to the context of their underlying information servers. Recently, this idea has been extended in the HyPersuit system [Wie96] by generalizing collections so that they may span documents from various servers.

The Harvest information discovery and access system [Bow95] provides an integrated set of tools for gathering information from diverse Internet servers. 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.

Subject gateways

A subject gateway, in network-based information access, is defined as a facility that allows easier access to network-based information resources in a defined subject area [Kir98]. 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.

Advanced gateways provide facilities for enhanced searching. For example the Social Science Information Gateway (SOSIG) [SOS99], incorporates a thesaurus containing social science terminology. This gives users the option of generating alternative terms/keywords with which to search the resource catalog. Another example of an advanced subject gateway is the Organization of Medical Networked Information (OMNI) [OMN99] which allows users to access medical and health-related information. OMNI also facilitates searches across other databases of resources such as databases of dental resources.

The key difference between subject gateways and the popular Web search engines, e.g., Alta Vista [Alt99], lies in the way the indexing is performed. Alta Vista indexes individual pages and not resources. For example, a large document consisting of many Web pages hyperlinked together via a table of contents would be indexed in a random fashion. In contrast this subject gateways, such as OMNI, index at the resource level, thus, describing a resource composed of many Web pages in a much more coherent fashion.

Furthermore, a subject gateway has the 'luxury' of being able to focus on a specific subset of the information space. Usually an area for which some well defined thesaurus is available.

Federated digital libraries

The most important problem specific to digital libraries with spatial distribution is the federation problem: making distributed collections of heterogeneous documents appear to be a single (virtually) integrated collection. Each such federation may address a specific domain area, e.g., biomedicine, computer science, social sciences and so on. In such federated digital libraries (FDLs) the difficulty lies in transforming a federation of multiple semi-structured heterogeneous sources (which lack coherence) into a single logical source. Here, we are faced with at least two major technical challenges. Firstly, document handling is hard as there is a large number of documents with differing structure and differing terminology. Secondly, due to the large number and variety of documents available, unless classification schemes are employed – so that document sources can be indexed in different ways and different levels of detail – distributed searching cannot be feasible [Sch96].

Multi-database systems

Multi-database (or federated) systems have as their aim the ability to access multiple autonomous databases through querying. The emphasis is on integration and sharing of distributed information and not on information discovery. A particular database may choose to export parts of its schema which are registered in a federal dictionary. A requesting database consults the federal dictionary for existing databases and then imports schema elements that it requires. While this approach might be appealing for a small number of interconnected databases it is clearly not scalable. Locating the right information in a large unstructured network of data dictionaries is extremely cumbersome, has limited potential for success and, more importantly, is error prone as it does not deal with terminology nuances.

More recently several research activities in the area have concentrated on the issue of creating semantically enhanced federated database dictionaries [BHP94], [Are93], [MS95], [CDA97]. Construction of conceptual ontologies on the basis of domain-specific terminologies and formalisms that can be mapped to description logics are also discussed in [KS97]. Some of the issues relating to the identification of semantically related information can be found in [BHP94], where the authors describe an approach that relies on an abstract global data structure to match user terms to the semantically closest available system terms. Concepts grounded on a common dictionary are defined in a domain and schema elements from component databases are manually mapped to these concepts. More recently, a different approach is taken by [KM97] where a domain-specific classification scheme is built incrementally by considering one schema at a time and mapping its elements in a concept hierarchy. However, both these approaches tend to centralize the search within a single logical index thereby defeating scalability by introducing performance limitations for large networks.

DAI: MultiAgent Systems

Most of the work on software agent systems has concentrated on improving information discovery methods on the WWW and adopt them for use within cooperating agent configurations. The protocols of the WWW provide purely keyword based index services and look up in collections of documents. Most DAI approaches employ some form of knowledge representation to enable more sophisticated representation of information sources and inferencing abilities [ONL94], [FEFP95], [BH95]. Two of the most notable activities which are elated to this work are: *information matchmaking* [KH95] and *information brokering* using *context logic* [FEFP95].

Matchmaking is an automated process whereby information providers and consumers are cooperating assisted by an intelligent facilitator utilizing a knowledge sharing infrastructure. Matchmaking depends on messaging and content languages and allows information providers and consumers to continuously issue and retract advertisements and requests, so that information does not become stale. This is particularly critical where information changes rapidly.

Fikes et al. [FEFP95] describe a tool-kit for information broker development based on the *Ontolingua* system [Gru92, Gru93]. Ontolingua is an integrated tool system for developing domain-specific ontologies in the Knowledge Exchange Format (KIF) and for translating the resulting ontologies into application-oriented representation languages. Their information brokers maintain declarative, logic-based, object-oriented models of their domain of expertise and the domains of expertise of their underlying resources.

3 Logical architecture of a multi-database information space

In order to improve efficient searching and gathering of schema information in large multi-database networks, the first task is to partition the multi-database information space into distinct, domain-specific, categories that are meaningful to database users. These categories can be formed by using some form of topic/subject based classification mechanism. Such classification mechanisms are common practices in library and information sciences, e.g., the INSPEC indexing and abstracting service covering most of the research literature in Computer Science and Electrical Engineering [Sch96]. Using domain-specific classifications to create logical clusters of databases makes searches more directed, meaningful and efficient. In addition, a directory of topics created as a result of domain-specific database categorization can also provide topic-specific searches and useful browsable organization of inter-component database schema information.

There are three basic principles that a system must address to allow for scalable information searching and gathering. Firstly, some organization of the underlying databases is needed to enable the discovery of data inter-relationships. Topic classification schemes are used for this purpose, as they summarize related information subspaces together. Secondly, this organizational structure must itself be scalable. In other words, both interactions with the resulting structure, as well as maintenance of the structure, must be scalable. Thirdly, users must be presented with a collection of tools (lexicographic, and user friendly graphical interfaces) which allows for easy exploration and interpretation of the information contents of the system. In the following, we address these issues in the context of a logical architecture for a multi-databases information space.

The logical architecture presented below has been among other inspired by ideas from the field of information retrieval, where documents are clustered to form a multi-layered information space [ACG91, AGM92, AMC95] that can be navigated by users in search of relevant documents.

3.1 Basic building blocks for information elicitation

Our approach to information searching and gathering in large database networks relies on logically partitioning a collection of networked databases into distinct topic based categories that are meaningful to users. This occurs by creating abstract representations of these topics as logical objects (henceforth referred to as topics). Database-content clustering algorithms can then be employed to automatically compute sets of related component databases – via their exported meta-data terms – and associate them with an appropriate topic, see Figure 1. The abstract objects representing these topics essentially represent centroids around which databases cluster, and are engineered to describe a particular domain. It is expected that the databases in a multi-database network target specific narrow domains, such as Geophysics, Biomedicine, Economics, Chemical Engineering and so on.

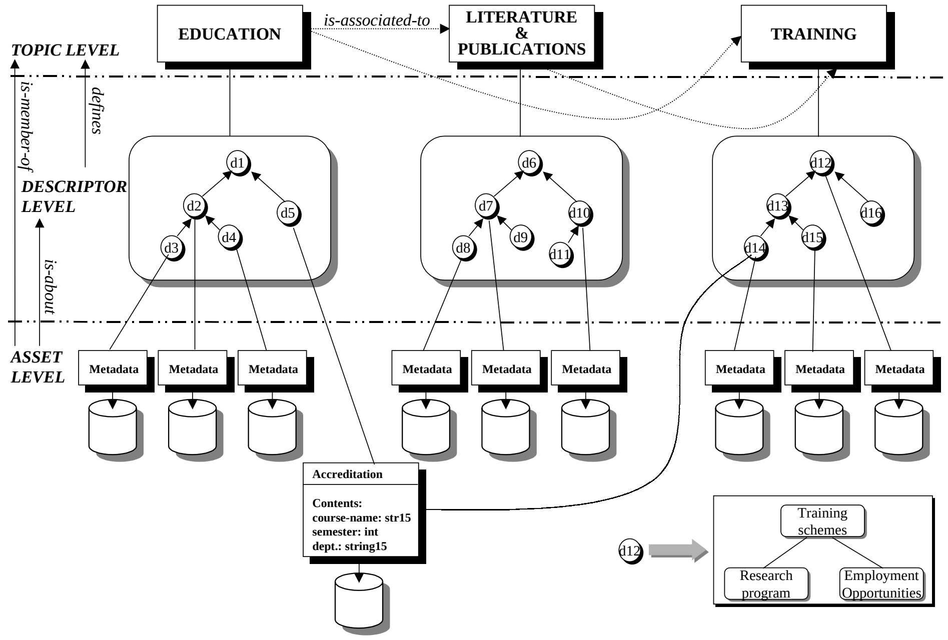
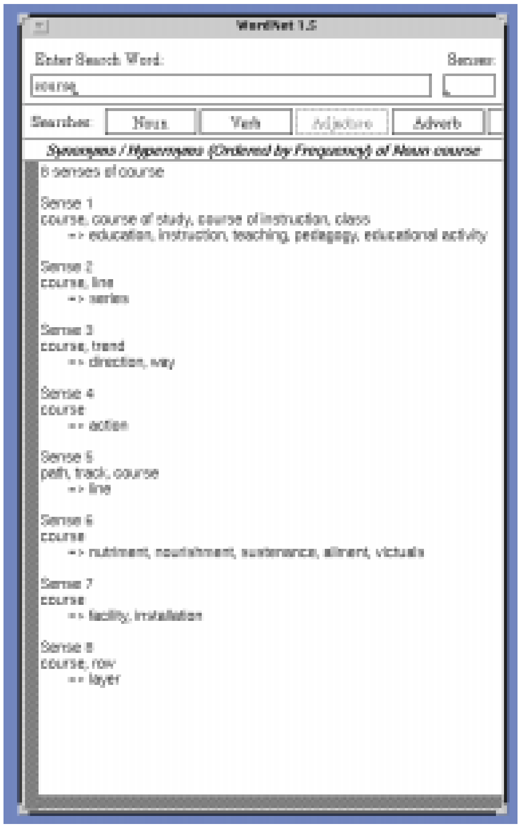


Figure 1: Three level organization of the information space.

To put the organization of a topic-based multi-database information space into perspective, we use a comprehensive example from an *Education & Training* multi-database network. This network connects educational and training service provider, publication provider, accreditation, and government agency database servers. This situation is shown in Figure 1 which provides a conceptually holistic view and cross-correlates information from the multiple database servers (referred to as assets). We will describe this process in two broad steps.

Firstly, we employ a *meta-data schema* to describe the structure and contents of each individual exported asset schema. Subsequently, distinct sets of asset schemas and their terms are logically aggregated to describe a particular subtopic. For example, aggregation of meta-data schemas which abstract assets containing information about courses, committees, accreditation-processes and so on, may represent subtopic such as a Accreditation. Subtopic terms are represented as composite objects in the individual meta-data schemas and may in their turn contain attributes. For instance, attributes such as course-name, credit-points, duration, etc would be contained in a course object which is part of the Accreditation subtopic. Meta-data schemas from each asset are organized in the form of graphs, which are called context graphs, see section 4. Each such graph correspond to a single meta-data schema while the aggregation of semantically related graphs may target a specific subtopic. Semantically related subtopics such as for example Accreditation and Enrollment-Program are also connected into a higher-level construct which we call a *topic*, e.g., Education, see Figure 1. Topics thus represent semantically related database 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 assets. The area of interest of a topic is defined in terms of a set of *descriptors*. Each descriptor is defined in terms of a context graph. An example can be found in Figure 1, where the context graph for descriptor d12 is depicted synoptically.

Secondly, to circumvent terminology mismatches and semantic drifts between disparate schema terms, topical synoptic knowledge and a standard vocabulary for term suggestions is supported by each topic. A topic materializes a class hierarchy depicting all terms within the topic sampled by the topic, e.g., Education. Each topic is characterized by its name and the context of its terms (term hierarchy and term descriptions) for each specific subject. Terms within a topic are shown to have a distinct meaning (sense) and context. The terminology context of a topic is usually provided by a standard ontology. An ontology can be defined as a linguistic representation of a conceptualization of some domain of knowledge [Gru93]. This ontology consists of abstract descriptions of classes of objects in a vertical domain, relationships between these classes, terminology descriptions and other domain specific information and establishes a common vocabulary for interacting with between the various information sources underlying a topic. Currently, WordNet [MRF^{+}90] is used to derive such a standard vocabulary for each of the topics. Hence, the topic structure is akin to an associative thesaurus and on-line lexicon (created automatically for each topic category). Thesaurus-assisted explanations created for each topic-based information subspace serve as a means of disambiguating term meanings, and addressing terminology and semantic problems. A topic is thus a form of a logical object (a kind of *a contextualized abstract view* over the content of large semantically related database collections) whose purpose is to cross-correlate, collate, and summarize the meta-data descriptions of semantically related network-accessible data[1], and thus it is grounded on a common standard ontology.



Enter Search Word: course

Synonyms / Hypernyms (Ordered by Frequency) of Noun course

```
8 senses of course

Sense 1
course, course of study, course of instruction, class
       => education, instruction, teaching, pedagogy, educational activity

Sense 2
course, line
       => series

Sense 3
course, trend
       => direction, way

Sense 4
course
       => action

Sense 5
path, track, course
       => line

Sense 6
course
       => nutriment, nourishment, sustenance, aliment, victuals

Sense 7
course
       => facility, installation

Sense 8
course, row
       => layer
```

Figure 2: Different senses of the term course.

Figure 2 shows an excerpt from WordNet showing eight different senses of Course. By identifying the sense which a topic about Courses is about, terminology use in the context of that topic may be standardized[2].

The topic-based multi-database configuration provides an appropriate frame of reference for both component database schema term indexing and user instigated searches. Figure 1 in particular illustrates that the *Education & Training* multi-database network comprises a set of topics such as

[1] Topics were termed "Generic concepts" or "Global concepts" in previous work [MBP95], [PM98].

[2] The WordNet lexicographic tool is presently used only for experimental purposes and will be replaced by an appropriate subject gateway in the near future.

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

A topically organized multi-database information space can be viewed as a Web-space, or a hypertext, that encompasses collections of exported meta-data ([ACG91, AGM92]). Such a multi-database information space partitions component databases into topically-coherent groups, and presents descriptive term summaries and an extended vocabulary of terms for searching and querying the vastly distributed information space of the component databases that underly it.

Assets in this network may become a member of more than one topic if they relate to their thematical foci. Individual topics are useful for browsing and searching large database collections because they organize information space in more comprehensible sub-spaces. For example, the Education topic provides a common terminology basis upon which database assets dealing with enrollments, courses, accreditation, etc, (see Figure 1), achieve knowledge of each others information content.

Although topics provide synoptic information about their underlying database clusters, *they do not require integration of the data sources.* This approach comes in strong contrast with approaches to semantic interoperability based on explicit integration of conceptual schemas on the basis of semantic lexica [BHP94], [CDA97]. The advantage of forming conceptual database clusters is that searches are goal-driven and the number of potential inter-database interactions is restricted substantially as it facilitates the distribution and balancing of resources via appropriate allocation to the various database partitions.

Formally, the basic building blocks of information space can be represented by three sets:

- let $\mathcal{TO}$ be the set of topics,
- let $\mathcal{DE}$ be a set of descriptors,
- let $\mathcal{AS}$ be the set of assets (databases) in information space.

Collectively, we refer to these elements as information objects:

$$\mathcal{IO} \triangleq \mathcal{AS} \cup \mathcal{DE} \cup \mathcal{TO}$$

As a first rule, the three base sets should not overlap, so the following axiom should apply:

[IS1] $\mathcal{AS}$, $\mathcal{DE}$, and $\mathcal{TO}$ are mutually disjoint sets.

Overall a multi-database network can be viewed in terms of three layers, see Figure 1. In this figure the $\mathcal{AS}$, $\mathcal{DE}$, and $\mathcal{TO}$ information objects each define a distinct level in the multi-database information space:

Topics layer A layer consisting of the topics that have been used to classify the assets (the underlying databases).

Descriptors layer A layer consisting of all descriptors used to provide a thematical description for the topics and define what the underlying assets (the databases) are about. Each descriptor is represented in terms of a context graph which organizes the terms contained in an asset. A formal description of descriptors can be found in section 3.3, while the process of creating and aggregating context graphs is described in section 4.

Asset layer A layer consisting of the actual databases; being the assets of information space.

3.2 Structure of information space

As shown in Figure 1, information space is characterized by a number of relationships between the constituting information objects in each layer. Let $\mathcal{RL}$ define the set of relationships in information space. For relationships, the following functions are presumed to be defined:

- $\mathsf{Src}, \mathsf{Dst} : \mathcal{RL} \rightarrow \mathcal{IO}$, defining the source and destination of a relationship.
- $\mathsf{Weight} : \mathcal{RL} \rightarrow \mathcal{WG}$, defining the certainty of a relationship. For weights, the existence of a combination operation $\otimes$ and a total order $<$ is presumed. For example, when using probabilities as weights, the combination operation may be computed as:

$$w \otimes v \triangleq 1 \Leftrightarrow (1 \Leftrightarrow w) \times (1 \Leftrightarrow v)$$

The relationships in information space should be irreflexive:

[IS2] $\forall_{r \in \mathcal{RL}} \left[\mathsf{Src}(r) \neq \mathsf{Dst}(r)\right]$

As an abbreviation we shall use:

$$x \rightarrow^{r}_{w} y \triangleq \mathsf{Src}(r) = x \wedge \mathsf{Dst}(r) = y \wedge \mathsf{Weight}(r) = w$$

When we are not interested in the specific weights associated to links, we will usually omit the weight w, and simply write $x \rightarrow^{r} y$.

Two key classes of relationships between information objects can be distinguished:

- *intra-layer relationships* ($\mathcal{IA} \subseteq \mathcal{RL}$), linking information objects of the same layer, viz. horizontal dimension, and
- *inter-layer relationships* ($\mathcal{IR} \subseteq \mathcal{RL}$), linking information objects of different layers, viz. vertical dimension.

These classes of relationships are presumed to form a partition of $\mathcal{RL}$:

[IS3] $\mathcal{IA}$ and $\mathcal{IR}$ are a partition of $\mathcal{RL}$.

The first class of relationships between information objects we focus on are the inter-layer relationships, i.e., those relationships that bridge information objects in the three layers. The following three classes of relationships exist between the information objects of the different layers:

- $\mathcal{AR} \subseteq \mathcal{IR}$ a set of aboutness relationships defining what the underlying assets are about,
- $\mathcal{DR} \subseteq \mathcal{IR}$ a set of defining relationships expressing the thematical focus of a topic in terms of descriptors, and
- $\mathcal{MR} \subseteq \mathcal{IR}$ a set of membership relationships identifying to which topics a given asset belongs.

The aboutness relationships in $\mathcal{AR}$ define what the underlying databases are really about in terms of the descriptors. In other words, the aboutness relationship defines the thematical scope of the databases

These classes of relationships are presumed to form a partition of $\mathcal{IR}$:

[IS4] $\mathcal{MR}$, $\mathcal{DR}$ and $\mathcal{AR}$ are a partition of $\mathcal{IR}$.

For the three classes of inter-layer relationships, the following predicates may be defined:

$$\begin{aligned} x \, \mathsf{IsMemberOf}_w \, y &\triangleq \exists_{r \in \mathcal{MR}} \, [x \rightarrow^r_w y] \\ x \, \mathsf{IsAbout}_w \, y &\triangleq \exists_{r \in \mathcal{AR}} \, [x \rightarrow^r_w y] \\ x \, \mathsf{Defines}_w \, y &\triangleq \exists_{r \in \mathcal{DR}} \, [x \rightarrow^r_w y] \end{aligned}$$

These relationships should indeed bridge the appropriate layers:

[**IS5**] $x \, \mathsf{IsMemberOf} \, y \Rightarrow x \in \mathcal{AS} \wedge y \in \mathcal{TO}$

[**IS6**] $x \, \mathsf{IsAbout} \, y \Rightarrow x \in \mathcal{AS} \wedge y \in \mathcal{DE}$

[**IS7**] $x \, \mathsf{Defines} \, y \Rightarrow x \in \mathcal{DE} \wedge y \in \mathcal{TO}$

The second class of relationships between information objects are concerned with intra-layer relationships. Many different types of relationships between information objects within a single layer may exist. For example:

- different types of associations,
- different types of part-whole relationships.

Instead of introducing a whole plethora of possible relationship types, we only focus on the following two general classes of relationships which are fairly representative:

- $\mathcal{AR} \subseteq \mathcal{IA}$ a set of associations,
- $\mathcal{PR} \subseteq \mathcal{IA}$ a set of part-whole relationships.

The membership relationship between assets and topics is fully derivable from the $\mathsf{IsAbout}$ and $\mathsf{Defines}$ relationships:

[**IS8**] $a \, \mathsf{IsAbout}_{w_1} \, d \wedge d \, \mathsf{Defines}_{w_2} \, t \Leftrightarrow a \, \mathsf{IsMemberOf}_{w_1 \otimes w_2} \, d$

Each asset must be about some descriptor with the maximum weight, e.g., 10/10, associated:

[**IS9**] $\forall_{a \in \mathcal{AS}} \exists_d \, [a \, \mathsf{IsAbout}_{1_{\mathcal{WG}}} \, d]$

where $1_{\mathcal{WG}}$ denotes the maximum weight from $\mathcal{WG}$ based on the total order. Each topic must be defined using a descriptor with the maximum weight associated:

[**IS10**] $\forall_{t \in \mathcal{TO}} \exists_d \, [d \, \mathsf{Defines}_{1_{\mathcal{WG}}} \, t]$

These classes of relationships are presumed to form a partition of $\mathcal{IA}$:

[**IS11**] $\mathcal{AR}$ and $\mathcal{PR}$ are a partition of $\mathcal{IA}$.

For the two general classes of intra-layer relationships, the following predicates may be defined:

$$\begin{aligned} x \, \mathsf{IsAssocTo}_w \, y &\triangleq \exists_{r \in \mathcal{AR}} \, [x \rightarrow^r_w y] \\ x \, \mathsf{IsPartOf}_w \, y &\triangleq \exists_{r \in \mathcal{PR}} \, [x \rightarrow^r_w y] \end{aligned}$$

Intra-layer relationships should indeed be *intra* layer:

[IS12] For each $S \in \{\mathcal{AS}, \mathcal{DE}, \mathcal{TO}\}$:

$$x \in S \wedge x \mathsf{\ IsAssocTo\ } y \Rightarrow y \in S$$

[IS13] For each $S \in \{\mathcal{AS}, \mathcal{DE}, \mathcal{TO}\}$:

$$x \in S \wedge x \mathsf{\ IsPartOf\ } y \Rightarrow y \in S$$

Part-of relationships cannot be weighted. In other words, they should always be of the maximum weight:

[IS14] $x \mathsf{\ IsPartOf}_w\ y \Rightarrow w = 1_{\mathcal{WG}}$

In the context of multi-database systems it does not make sense to allow for subset relationships between topics as topics are autonomous and disjoint from each other. In other words, we have:

[IS15] $x \mathsf{\ IsPartOf}_w\ y \Rightarrow x, y \notin \mathcal{TO}$

The part-of relationship is transitive and irreflexive:

[IS16] $x \mathsf{\ IsPartOf\ } y \mathsf{\ IsPartOf\ } z \Rightarrow x \mathsf{\ IsPartOf\ } z$

[IS17] $\neg(x \mathsf{\ IsPartOf\ } x)$

In the context of multi-database systems, it furthermore does not make sense to introduce intra-layer relationships between the underlying databases due to the unnecessary complexity and also because all databases are strongly related to their encompassing topic.

In other words, in this context we have:

[IS18] $x \mathsf{\ IsAssocTo\ } y \Rightarrow x, y \notin \mathcal{AS}$

[IS19] $x \mathsf{\ IsPartOf\ } y \Rightarrow x, y \notin \mathcal{AS}$

In future research we will apply the results as presented in this paper to more general forms of assets, such as web-pages, documents, etc. In these latter cases it is indeed sensible to cater for associative and part-of relationships between the assets. For example, documents referring to each other imply an associative relationship while a chapter is-part-of a book.

The set of descriptors that is associated to a topic by the Defines relationship is essentially the thematic scope of a topic:

$$\mathsf{Thema}(t) \triangleq \{d \mid d \mathsf{\ Defines\ } t\}$$

The theme of a topic should have a unique top element based on the IsPartOf relationship on descriptors:

[IS20] $\forall_{t \in \mathcal{TO}} \exists!_{x \in \mathsf{Thema}(t)} \left[\neg \exists_{y \in \mathsf{Thema}(t)} [x \mathsf{\ IsPartOf\ } y]\right]$

If two topics are associated to each other with some weight, then this association should somehow be reflected by the associations between descriptors associated to the respective topics:

[IS21] $t_1 \mathsf{\ IsAssocTo}_w\ t_2 \Rightarrow \quad w = \bigotimes_{r \in B(t_1, t_2)} \mathsf{Weight}(r)$, where:

$$B(t_1, t_2) \triangleq \left\{r \in \mathcal{AR} \,\middle|\, \exists_{d_1 \in \mathsf{Thema}(t_1), d_2 \in \mathsf{Thema}(t_2)} [d_1 \rightarrow^r d_2]\right\}$$

The set of descriptors used to express what an asset is about, is referred to as the *aboutness* of that asset:

$$\mathsf{Aboutness}(a) \triangleq \left\{ d \mid a \,\mathsf{IsAbout}\, d \right\}$$

The aboutness of an asset should have a unique top element as well:

[IS22] $\forall_{a \in \mathcal{AS}} \exists!_{x \in \mathsf{Aboutness}(a)} \left[\neg \exists_{y \in \mathsf{Aboutness}(a)} \left[x \,\mathsf{IsPartOf}\, y \right] \right]$

The aboutness of assets is a rather complex relationship. In [BL97, PB99] detailed studies have been made of what aboutness really is. In these publications several rules may be found governing the aboutness relationship. An example of such a rule would be:

[IS23] $x \,\mathsf{IsPartOf}_{w_1}\, y \wedge x \,\mathsf{IsAbout}_{w_2}\, c \Rightarrow y \,\mathsf{IsAbout}_{w_1 \otimes w_2}\, c$

For more detailed rules on aboutness, refer to e.g. [BL97, PB99].

3.3 Descriptors

The descriptors defining the theme of a topic play a very important role in information space. It provides users, as well as the system, with a description of a context in terms of some suitable language. Topics in information space may be described using such mechanisms as: keywords, index expressions [Cra78], noun phrases, or different forms of conceptual graphs [Sow84]. Most readers will be familiar with keywords to express what a text database or document is about. The language of index expressions was introduced as a description language that would allow keywords to be put in relation to each other. For example:

$$\mathsf{polution\ of\ rivers\ in\ Europe} \quad \text{or} \quad \mathsf{economic\ benefits\ of\ Euro}$$

rather than simply:

$$\{\mathsf{polution}, \mathsf{rivers}, \mathsf{Europe}\}$$

The effectiveness of index expressions in an information retrieval context has been studied and reported in [Bru90]. The language of index expressions is actually a subset of the language of noun-phrases. The use of conceptual graph like languages for retrieval purposes has been studied in e.g. [ACG89, Mya92, Bra98]. These publications point out that it is useful to separate concepts and their description as a specific concept may quite well be represented in different languages using a specific description for each of the languages.

Thus far, this paper has treated descriptors abstractly. In this subsection the descriptors are defined in more detail. As keywords and index expressions can both be represented as conceptual graphs, we base the formal definition of descriptors on conceptual graphs.

To define descriptors formally, we start out from two base sets:

- a set $\mathcal{CO}$ of concepts, and
- a set $\mathcal{ED}$ of edges.

In addition, the following functions are needed:

- $\mathsf{Src}, \mathsf{Dst} : \mathcal{ED} \rightarrow \mathcal{CO}$, defining the source and destination of an edge.
- $\mathsf{Name} : (\mathcal{ED} \cup \mathcal{CO}) \rightarrow \mathtt{String}$, the name of a relationship or a concept.

  The names of the relationships could, for example, be based on the ones as pre-defined by WordNet, such as: hypernyms, hyponyms, part-of, and pertains-to.

- $\mathsf{Explain} : \mathcal{CO} \rightarrow \mathtt{String}$, some additional explanation of the concept.

Together $\mathcal{CO}$, $\mathcal{ED}$, Src, Dst, Name, and $\mathsf{Explain}$ define a *concept space*.

Using this concept space, the set of (possible) descriptors can be defined as follows:

$$\mathcal{DE} \triangleq \left\{ E \subseteq \mathcal{ED} \;\middle|\; \begin{array}{l} \text{The set of edges } E \text{ spans a connected subgraph and} \\ \forall_{c,d \in \mathsf{Concepts}(E)} \left[\mathsf{Name}(c) = \mathsf{Name}(d) \Rightarrow c = d\right] \end{array} \right\}$$

where $\mathsf{Concepts}(E) \triangleq \bigcup_{e \in E} \{\mathsf{Src}(e), \mathsf{Dst}(e)\}$. Note that within one descriptor, the names of concepts must be unique. For descriptors $d, e \in \mathcal{DE}$, the $\mathsf{IsPartOf}$ relationship can be defined formally as:

$$d \,\mathsf{IsPartOf}_{1_{\mathcal{WG}}}\, e \triangleq d \subseteq e$$

The $\mathsf{IsPartOf}$ relationship for the descriptors d and e can simply be defined formally as:

$$d \,\mathsf{IsPartOf}\, e \triangleq d \subseteq e$$

By using ontologies more contextual information may be added to the descriptions of concepts. For instance, in Figure 2, a WordNet [MRF^{+}90] style context for the Course concept is shown. There it is shown what particular meaning (sense) and context of the word Course is the focus of the concept. As mentioned before, two different concepts (in different topics) may indeed have the same name, while they are concerned with different senses of the same name/term.

3.4 Meta-data

To each of the information object, so-called meta-data attributes may be associated [WGMD95]. Such meta-data attributes may for example be concerned with:

> *authorship*, *date of creation*, *medium*, *file format*, *pricing*, *quality*, and *location*.

Meanwhile different emerging metadata standards have come into being. For example, Dublin Core [Dub99] or PICS [PIC99]. Meta-data attributes may be formalized as follows:

- $\mathsf{Name} : \mathcal{IO} \rightarrow \mathtt{Names}$ providing the name of the information object.
- $\mathsf{Location} : \mathcal{AS} \rightarrow \mathtt{URI}$ yielding the physical location/address of an asset (database). For example, in terms of a URI (a Universal Resource Indicator, such as the well-known URL from the World-Wide-Web).
- $\mathsf{Layer} : \mathcal{IO} \rightarrow \{\mathsf{Topics}, \mathsf{Descriptors}, \mathsf{Assets}\}$ returning the name of the layer an information object belongs to.

For Layer we obviously have:

[IS24] $\forall_{t \in \mathcal{TO}} [\mathsf{Layer}(t) = \mathsf{Topics}\ ] \wedge \forall_{d \in \mathcal{DE}} [\mathsf{Layer}(c) = \mathsf{Concepts}] \wedge \forall_{a \in \mathcal{AS}} [\mathsf{Layer}(a) = \mathsf{Assets}\ ]$

4 Clustering of Databases

In the following we describe a general methodology that aids in clustering databases and creating their corresponding topic nodes in information space. Key criteria that have guided this methodology are: scalability, design simplicity and easy to use structuring mechanisms.

4.1 Describing a database

In order to initially cluster component databases, a high level description of the database contents in the form of a descriptor must first be derived.

To demonstrate this consider the example of the Universal_Accreditation database, which deals with academic institutions and accreditation processes and is connected to the *Education & Training* multi-database network. This database contains entities such as courses, committees, accreditation, processes, etc. In order to become part of a topic-based multi-database network the schema terms of this database are represented in the form of a *context graph*, which is essentially a form of a *conceptual graph*, is created that interconnects terms (concepts) on basis of their their semantic relatedness. To achieve this we use a variant of an information retrieval technique called, the *star technique*. With this technique, a concept is selected and then all concepts related to it are placed in a class [Kow97]. Concepts not yet in a class are selected as new seeds until all concepts are assigned to a class. The variant of the star technique that we are using starts with a concept, then an additional concept that is related to an already selected concept is represented as a another class and is connected to the selected concept. The new concept is then selected as a pivot and the process is repeated until no new concepts can be added. For example, the context graph for the Universal_Accreditation asset, as depicted in Figure 3 contains nodes which correspond to the committee, institutions, courses, etc.

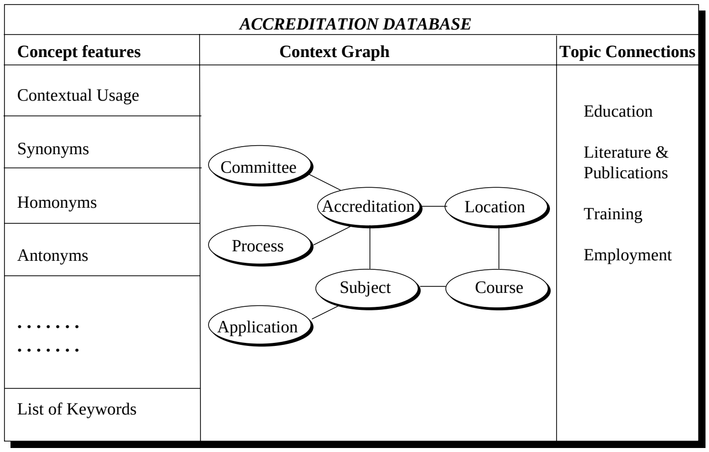


Figure 3: Describing a component database.

The context graph edges depict inter–connections (association, generalization, specialization or containment) between the concepts within a particular asset. Concept interrelations are determined on the basis of a reference lexicographic substrate that underlies the concepts in information space. For this purpose, as already explained, we use the lexicographic system WordNet [MRF+90] that supports semantic concept matching through the use of an extensive network of word meanings connected by a variety of textual and semantic relations.

If d is the context graph (descriptor) that is derived for a database (asset) a using the algorithm as sketched above, then we know: $a\,\mathsf{IsAbout}\,d$. Furthermore, using the IsPartOf relationship for descriptors, we can also infer that for any context graph e, if $e\,\mathsf{IsPartOf}\,d$ we have $a\,\mathsf{IsAbout}\,e$.

To facilitate clustering and discovery of information, we require that an asset (e.g., Universal_Accreditation) can be totally described in terms of three sections which contain a synoptic description of the meta-data content of the asset; associations between meta-data terms in the form of a semantic-net; and finally, links from these descriptions to other related assets in the network. This information can be viewed by users of the system once they have chosen a component database that potentially matches their interests (see section 5).

Figure 3 illustrates that each database node contains the following sections:

- a *concept features*,
- a *context graph*, and
- a *topic connections* section.

The *concept features* section contains additional information concerning the concepts used in the context graph, essentially providing the Explain function for concepts in the context graph. These descriptions include abstract descriptions of terms in the domain such as their sense (unique), relationships between these terms, composition of terms, terminology descriptions, hypernym, hyponym, antonyms-of, part-of, member-of, pertains-to relations, contextual usage (narrative descriptions), a list of keywords, and other domain specific information, that apply to the entire collection of members of a topic. Moreover, it may include other useful details such as: geographical location of databases, access authorization and usage roles, explanations regarding corporate term usage and definitions, domains of applicability, charge costs, and so on. The feature descriptions entries are partially generated on the basis of WordNet and contain information in the form represented in Figures 2 and 4.

The *context graph* section contains a non–directed graph which connects the concepts as described by the *concept features* section. Except for viewing purposes when navigating information space, the context graph used in the clustering of databases to form topics. Each of the concept nodes defines (in conjunction with its respective entry in the feature descriptions window) a common structured vocabulary of terms and term relationships relevant to that concept. Finally, the topics connection section shows to which topics the Universal_Accreditation asset is related to in the network.

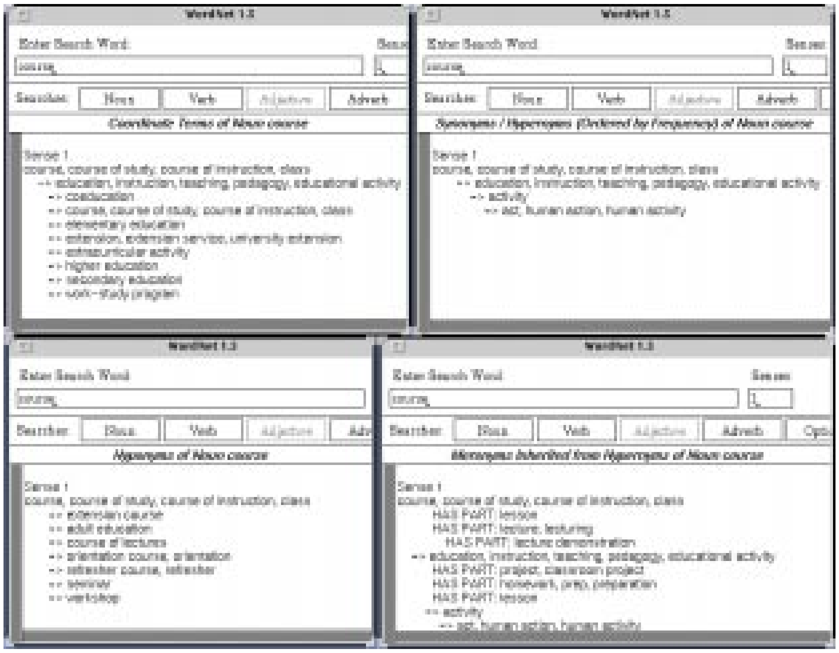
4.2 Similarity-based clustering of databases

Similarity-based clustering of database schemas organizes databases into related groups based on the concepts that are referred to by the *concept features* section of their database description (see Figure 3) they contain and the link structure of their context graphs.

Our clustering algorithm determines the similarity between two graphs (representing two different database schema meta-data) based on both concept similarity and link similarity factors. This is accomplished in two steps. 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. We chose this concept/link similarity-based algorithm because it is relatively easy to implement and avoids generating very large clusters.

Concept-based similarity: this is calculated using cluster analysis techniques [Eve81] to identify co–occurrence probabilities – representing the degree of similarity – between two discrete concepts.

Our similarity metric is based on the meaning of the collection of terms representing the terminological context (viz. semantic-levels) of a particular concept, e.g., Course, see Figure 2. The comparison is based on: a conversion of each context graph node Committee, Process,



Coordinate Terms of Noun course

```
Sense 1
course, course of study, course of instruction, class
   => education, instruction, teaching, pedagogy, educational activity
       => coeducation
       => course, course of study, course of instruction, class
       => elementary education
       => extension, extension service, university extension
       => extracurricular activity
       => higher education
       => secondary education
       => work-study program
```

Synonyms / Hypernyms (Ordered by Frequency) of Noun course

```
Sense 1
course, course of study, course of instruction, class
       => education, instruction, teaching, pedagogy, educational activity
           => activity
               => act, human action, human activity
```

Hyponyms of Noun course

```
Sense 1
course, course of study, course of instruction, class
       => extension course
       => adult education
       => course of lectures
       => orientation course, orientation
       => refresher course, refresher
       => seminar
       => workshop
```

Meronyms Inherited from Hypernyms of Noun course

```
Sense 1
course, course of study, course of instruction, class
        HAS PART: lesson
        HAS PART: lecture, lecturing
           HAS PART: lecture demonstration
    => education, instruction, teaching, pedagogy, educational activity
        HAS PART: project, classroom project
        HAS PART: homework, prep, preparation
        HAS PART: lesson
      => activity
         => act, human action, human activity
```

Figure 4: Further WordNet details for course.

Subject, Course, etc. (see Figure 3) to a corresponding matrix of noun terms (containing the entire terminological context of a concept); and a subsequent comparison of terms within these matrixes.

A matrix a for concept c with m columns, and n rows, should be organized such that:

1. Each column contains synonyms describing the same concept:

$$\forall_{1\leq i<j\leq m; q\leq k\leq n}\,[a_{i,k} \text{ described the same concept as } a_{j,k}]$$

   For example, Course, Course-of-study, Course-of-lectures, etc.

2. There must be a unique column in which the first element corresponds to the name of the concept:

$$\exists!_i\,[\mathsf{Name}(c) = a_{i,1}]$$

3. The columns in the matrix should be ordered from general to more specific. In other words, terms to the left should (pairwise) not be more specific than terms to the right:

$$\forall_{1\leq i<j\leq m; 1\leq k\leq n}\,[a_{i,k} \text{ is not a more specific term than } a_{j,k}]$$

   For example, Education, Educational-activity, are more general terms than Course, while Computer-science-course is a more specific term than Course.

Similarity analysis is mainly based on statistical co–occurrences of terms based on techniques which have been successfully used for automatic thesaurus generation of textual databases [Eve81], [SB88]. In fact, we base our term-based similarity on the improved *cosine* formula [SB88] which is used to calculate the semantic distance between the vector for an item in a hierarchical thesaurus and the vector for a query item. To provide the right ontological context for semantic term matching, we use again the massive semantic net WordNet [MRF$^+$90].

Comparison of the conceptual structure of two context graphs: 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 area presented in [RB89]. 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. During this process similarity between nodes is established by considering the edges separating the nodes in the context graph as well as the actual graph structure. Some early results regarding the comparison and clustering process are described in [MMW96].

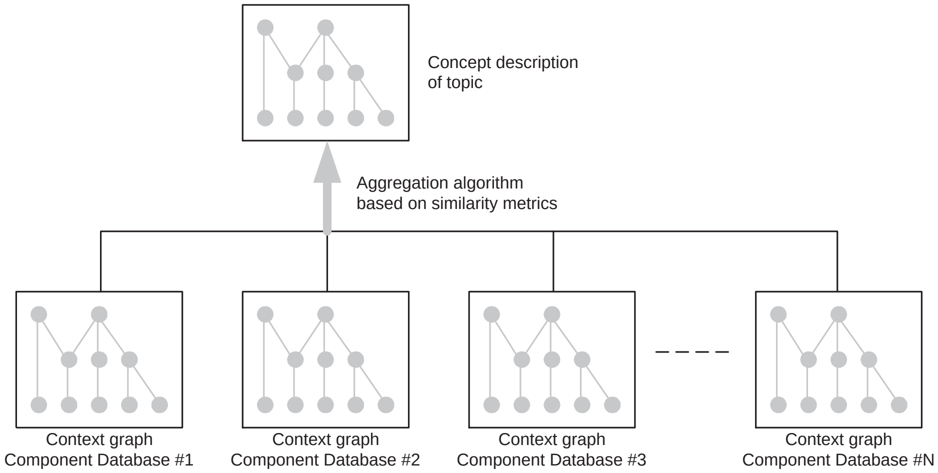


Figure 5: Clustering interrelated component schema terms.

Once similarity between the context graphs describing what the underlying databases are about has been established, the context graphs can be aggregated to create topics. The aggregation of the context graphs from various component databases, results in the clustering of inter–related database schema terms, see Figure 5. This aggregation is performed on the basis of the reference lexicographic substrate (WordNet). The aggregation algorithm employed does not integrate the aggregated databases, as is the usual case with other approaches [KS97], but rather links concepts at the topics level with corresponding concepts in its underlying cluster of database context graphs.

For each database cluster, a topic is created to represent the area of interest that the group embodies, e.g., a Education topic for the Accreditation, Tertiary Education and Enrollment Program databases, see Figure 1. The aggregated context graph is used as a base to define the thematic scope of the particular topic. Using the IsPartOf relationship on descriptors, this description can be completed with all sub-descriptors.

When clustering the databases, some cut-off value needs to be employed to express how similar two databases should be for them to be placed in the same cluster. This cut-off value then essentially expresses the internal *cohesion* of the resulting cluster (and topic).

4.3 Deriving sub-topics

When some topic t contains a large number of databases, it may be useful to split t into several smaller topics. This can be done by re-clustering the underlying set of databases using a higher cut-off value. Each of the resulting new topics has an internal cohesion that is higher than the original topic. If T is the set of resulting sub-topics, then these may be linker to the original (large) topic using the IsPartOf relationship:

$$s \in T \Rightarrow s \ \mathsf{IsPartOf}_{1_{\mathcal{WG}}} \ t$$

where $1_{\mathcal{WG}}$ represents the maximal value from Weight.

4.4 Interrelating topics

Once the topics have been created, it is also possible to use the similarity algorithm to compute the similarity between the (aggregated) context graphs associated to topics. This similarity can then be interpreted as the similarity between topics. Using some cut-off value, the similarity relationship between topics can be used to fill the IsAssocTo relationship between topics. The similarity weight can be used as a weight on these relationships. In other words, if c is the cut-off value, and $\mathsf{Sim}(t_1, t_2)$ expresses the similarity between two topics, we have:

$$\mathsf{Sim}(t_1, t_2) = w \land w \geq c \Rightarrow t_1 \,\mathsf{IsAssocTo}_w\, t_2$$

5 Navigation and querying

Information elicitation spans a spectrum of activities ranging from a search for a specific data-item(s) (contained in possibly several component databases) to a non-specific desire to understand what information is available in these databases and the nature of this information. This section is concerned with navigating and querying information space as techniques for information elicitation.

5.1 Stratified architecture

As mentioned before, information space can be grouped into an assets, descriptors, and a topics layer. These three layers allows us to organize information space as a three-level stratified hypermedia as reported in e.g. [AAC$^+$89, BW90, ACG91]. Organizing a multi-database information space as a stratified hypermedia enables users to navigate information space in a natural way.

In [AAC$^+$89, BW90, ACG91] these ideas have been applied in the connect of documents, while in the work reported in [HPW96] these ideas are translated to query formulation on databases. The latter is in line with the application of stratified hypermedia as a way of navigating a multi-database information space.

In a stratified hypermedia, each layer consists of a set of nodes and links. For a set X of information objects in a particular layer, $\mathcal{AS}$, $\mathcal{DE}$, and $\mathcal{TO}$, the corresponding layer is defined by: $\left\langle X, \mathsf{IsAssocTo}^X, \mathsf{IsPartOf}^X \right\rangle$, where:

$$\begin{aligned} x \,\mathsf{IsAssocTo}^X\, y &\triangleq x, y \in X \land x \,\mathsf{IsAssocTo}\, y \\ x \,\mathsf{IsPartOf}^X\, y &\triangleq x, y \in X \land x \,\mathsf{IsPartOf}\, y \end{aligned}$$

The three layers of information space can then be defined as follows:

1. Assets layer: $\left\langle \mathcal{AS}, \mathsf{IsAssocTo}^{\mathcal{AS}}, \mathsf{IsPartOf}^{\mathcal{AS}} \right\rangle$.
2. Descriptors layer: $\left\langle \mathcal{DE}, \mathsf{IsAssocTo}^{\mathcal{DE}}, \mathsf{IsPartOf}^{\mathcal{DE}} \right\rangle$.
3. Topics layer: $\left\langle \mathcal{TO}, \mathsf{IsAssocTo}^{\mathcal{TO}}, \mathsf{IsPartOf}^{\mathcal{TO}} \right\rangle$.

The above definitions are general and can be used for all sorts of information assets, e.g., documents. However, in the case of multi-database systems for $\mathcal{AS}$ both relationships $\mathsf{IsAssocTo}^{\mathcal{AS}}$, $\mathsf{IsPartOf}^{\mathcal{AS}}$ are empty whereas for the $\mathcal{TO}$ layer the $\mathsf{IsPartOf}^{\mathcal{TO}}$ is empty, see the corresponding axioms in section 3.

As $\mathcal{AS}$, $\mathcal{DE}$, and $\mathcal{TO}$ are disjoint sets, these layers are disjoint as well. When navigating a stratified hypermedia architecture, users may not only want to navigate within a single layer, they will also want to navigate between layers.

The $\mathsf{JumpsTo}$ relationship is a generalization of $\mathsf{IsMemberOf}$, $\mathsf{Defines}$ and $\mathsf{IsAbout}$ and is used for inter-layer navigation. This generalization is defined as follows:

$$x\,\mathsf{JumpsTo}\,y \;\triangleq\; \exists_{r\in\mathcal{IR}}\left[\{x,y\}=\{\mathsf{Src}(r),\mathsf{Dst}(r)\}\right]$$

This three-tier architecture is the key ingredient to finding information in distributed, scalable systems. It generates a semantic hierarchy for database schema terms in layers of increasing semantic detail. Most searches will initially target the richest semantic level, viz. the topics layer, and percolate via the descriptors layer to the assets layer in order to provide access to the contents of an asset. This type of content-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 methodology results in a simplification of the way that information pertaining to a large number of interrelated database schemas can be viewed and more importantly it achieves a form of global visibility [Pap95].

5.2 Node presentation

When navigating the stratified hypermedia as derived from an information space, users travel from information object to information object via intra-layer or inter-layer links. Each information object is presented to a user by providing three types of information to the user:

- the current location in information space,
- the possible steps to continue the journey in information space,
- more information about the current location.

Let f be the information object in information space where the user is currently at, then the above three types of information can be further specialized as:

- The description of current location consists of $\mathsf{Name}(f)$, $\mathsf{Layer}(f)$, and $\mathsf{Size}(f)$.
- The possible continuations consists of six sets:
  1. Refinements: $\left\{x \;\middle|\; x\,\mathsf{IsPartOf}\,f\right\}$
  2. Enlargements: $\left\{x \;\middle|\; f\,\mathsf{IsPartOf}\,x\right\}$
  3. Associations: $\left\{x \;\middle|\; f\,\mathsf{IsAssocTo}\,x\right\}$
  4. Jumps to the topics layer: $\left\{x\in\mathcal{TO} \;\middle|\; f\,\mathsf{JumpsTo}\,x\right\}$
  5. Jumps to the descriptors layer: $\left\{x\in\mathcal{DE} \;\middle|\; f\,\mathsf{JumpsTo}\,x\right\}$
  6. Jumps to the assets layer: $\left\{x\in\mathcal{AS} \;\middle|\; f\,\mathsf{JumpsTo}\,x\right\}$

  In describing the possible continuations to users, the system may choose to use $\mathsf{Name}(x)$. In the case of descriptors, it may be useful to also include some synoptic version of the context graph.

- Other information about the current focus f may consist of any other meta-data available on f.

In presenting this information to a user, the system may opt to use multiple screens. For example, Figure 6 illustrates how a user moves from the topic to the descriptor and the asset layer when looking for information relating to Education. Here we assume that the user started from the Education topic layer and that the concepts within the descriptor layer are organized in ascending order of specificity. For instance from activity which is a highly abstract term, to terms such as course of study, course of lectures, and so on. We can move between these concepts by using the refinement field (operation). In Figure 6 we assume that the user was interesting in finding about assets dealing with Accreditation.

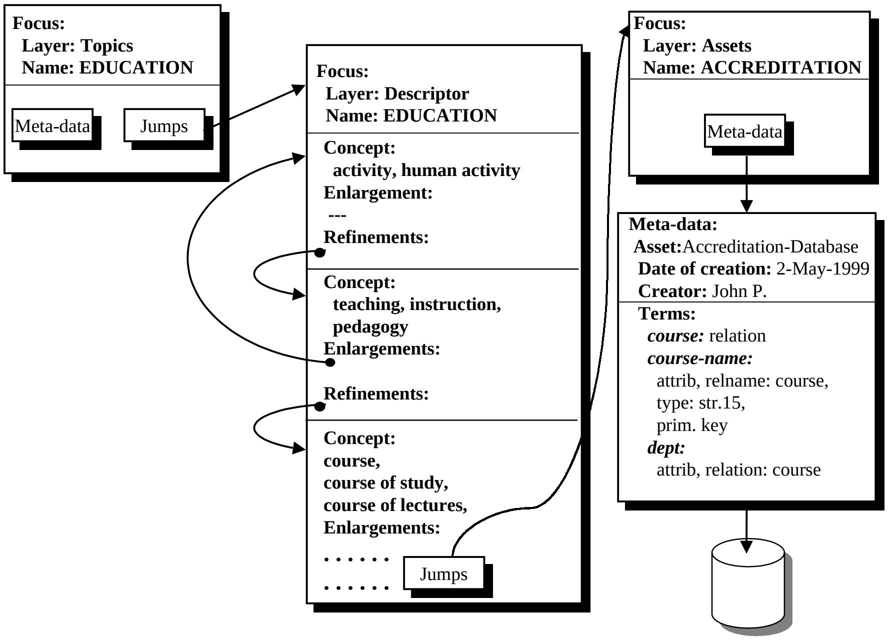


Figure 6: Screens representing information objects.

The order in which the continuation options are listed in the nodes should ideally be determined by the relevance of these options to the user. For example, in [BW95] a mechanism is discussed by which the relevance of these options can be derived based on a user's search behavior thus far. This may be combined with the relative weights that can be associated to the links in information space.

5.3 Navigation Techniques

There are two basic modes in which searching of the system may be organized. These search modes depend upon the nature of the information a user is attempting to access, and how this information relates to the database that the user is operating from. Serendipity, exploration and contextualization are supported by navigating the stratified hypermedia spanned by information space. In such cases the user is interested in finding out about a particular topic rather than a specific information (schema) item. We call this former form of exploration *index-driven.*

Alternatively, if a user is seeking data which is closely related or allied to her/his local database, then searching may be organized around the links of this database to other topics in information space. We refer to this form of exploration as *concept-driven.* Concept-driven querying is the subject of a previous publication [PM96]. In this paper, we are mainly concerned with index-driven searches.

Index-driven navigation allows the users to deal with a controlled amount of material at a time, while providing more detail as the user looks more closely. This form of searching is related to the dynamic indexing schemes and incremental discovery of information requirements for information elicitation. An index-driven search will usually start out from some name $n \in \texttt{Names}$ of a concept. From this point, the search will gradually percolate down to the required level of specificity. This process starts by treating n as a descriptor and matching this to the set of available concepts. This results in a set ($D(n)$ say) of potentially interesting descriptors. This set may become rather large. It may therefore be sensible to limit this set to the elements that are lowest (i.e. least specific) in the $\mathsf{IsPartOf}$ hierarchy:

$$\left\{d \in D(n) \;\middle|\; \neg \exists_{e \in D(n)} \left[e \;\mathsf{IsPartOf}\; d\right]\right\}$$

By presenting the resulting set of concepts, users can consequently home in on the intended interpretation of n, and consequently refine their focus of interest by navigating through the descriptors layer.

If a user is already focussed on a specific topic (t say), the set $D(n)$ can be limited in advance to the set $\mathsf{Thema}(t) \cap D(n)$.

5.4 Querying of Domain Meta-Data

When the user needs to further explore the search target, *intensional*, or schema queries [Pap95] – which return meta-data terms from selected schema terms – can be posed to further restrict the information space and clarify the meaning of the information items under exploration. Such domain-specific queries should not be confused with queries which target the data content of the assets (to which we refer to as *distributed* queries/transactions). Intensional queries are particularly useful for assisting users who are unfamiliar with the vocabulary of terms that can be used in connection with distributed queries/transactions or with the range of information that is available for responding to distributed queries. Sample intensional queries related to the descriptions as depicted in Figure 2 may include the following:

query-1: *Find the set of common super-terms of course.*

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

query-3: *Find the smallest common super-term of course of lectures and workshop.*

query-4: *Find all parts of the term course.*

query-5: *Which are the common properties of refresher course and seminar?*

query-6: *Find all terms which contain the properties lesson and classroom project.*

query-7: *What is the definition of the term refresher course?*

All of the above queries - except for the last one - are rather intuitive. The last query returns a narrative description of the requested term in English (if available).

Finally, when users feel sufficiently informed about the contents and structure of asset schema terms they have explored, they can pose meaningful distributed database requests which target the data content of the relevant component databases.

6 Experimentation

The framework that we described in this paper is being implemented on Sun SparcStations under Solaris 2 using GNU C++ and CGI scripts. In order to evaluate automated clustering a test platform based on the clustering of about hundred networked databases has been created. There are two basic areas of experimentation being pursued. Firstly, there is the question of how well the initial automatic clustering of databases based on each asset's meta-data description can be performed. That is, the scalability question of finding appropriate initial relationships in the presence of large numbers of information sources. The types of experiments performed here are somewhat allied with the field of information retrieval and clustering. The second set of experiments, on the other hand, deals with the processing and communications necessary to support the underlying distributed structure by which the generic concepts and their inter-relationships are implemented, queried and updated. This second group of experiments thus has its roots in the fields of distributed/parallel processing and communications performance.

In a similar vein to IR experiments, the first set of experiments are based on the notion of retrieval and accuracy (as defined within IR). To achieve this, a collection of a hundred relational databases has been procured from a large organization's collection of information systems. A manual clustering of these was then performed by a domain "expert" who had full intimate knowledge of the organization's environment. This clustering was essentially based on where each database fitted into the various departments within the organization, and how these departments interacted/overlapped – the latter being identified via analysis of database table usage within the various departments. Thus, we clustered assets based on the actual usage of data from the various information components as dictated by the organization of the environment that the databases were set up to model in the first place – but in a macro (organization wide) sense rather than a micro (department based) sense.

Experiments have been performed (and continue to be performed) to:

1. identify if automatic clustering can achieve a "near perfect" initial organization of the database collection - or at least be statistically significantly better than "raw" automatic clustering, which involves the identification of an appropriate heuristic for measuring the similarity between database descriptions;

2. compare results against other standard automatic clustering packages (e.g., those found in IR);

3. determine what set of descriptive "primitives" are essential (and minimal) to achieve a satisfactory degree of clustering;

4. determine the "robustness" of the description process – i.e., give some indication of how much variation there can be within a description before the automatic clustering becomes unsatisfactory.

Currently, experiments have been performed using a "full" asset description involving the synonyms, generalizations and terms senses, as well as the structural relationships between these terms, see Figure 4. Initially, the term matching component was based on the standard similarity metric proposed by Dice [Eve81], and the structural similarity was based on the notion of spreading activation energy [MMW96]. It was found, however, that the accuracy and retrieval of this particular approach was not significantly better than the clustering of the "raw" database descriptions using Dice's method directly. Upon analysis it was discovered that performance was degraded due to the un-directed nature of the context graph. Thus, in a subsequent set of preliminary experiments, the notion of spreading activation energy was dropped, and a ranking of similarity based on the hierarchy of the graph was introduced. This resulted in a huge improvement in the retrieval and similarity figures which indicated the automatic clustering to be significantly better than the base-line clustering.

7 Summary and Future Work

The topic-based organization of a multi-database network supports semantic reconciliation of autonomous interconnected data sources as it helps the users understand what information is available through the network; helps them categorize and configure their information demands on the basis of the information available to them; and assists them to semantically disambiguate their specified terms against those provided by the database schemas in a multi-database network. This architecture enables users to gather and rearrange information from multiple networked databases in an intuitive and easily understandable manner. Large-scale searching is guided by a combination of lexical, structural and semantic aspects of schema terms in order to reveal more meaning both about the contents of a requested information item and about its placement within a given database context. Experience with this configuration suggests the clustering mechanisms used provide a valuable discovery service to end users, and that the logical organization used supports the ability of the system to scale with modest increases in topic label sizes.

Future work addresses the semi-automatic generation of link weights based on term co-occurrences using statistical/probabilistic algorithms. In IR these algorithms use word and/or phrase frequency to match queries with terms [Eve81]. In the current prototype link weights are established at a clustering phase on a tentative basis only. However, it is expected that during execution link weights to topics may need to be updated (strengthened or weakened) over time depending on interaction, new topics may be formed, and existing topics may need to merge. The next suite of experiments to be performed will deal with the characteristics of the link weight update and topic split/merge processes. From this policies will be developed (e.g. delayed/batch updating of topic information), and then evaluated.

References

[AAC$^+$89] M. Agosti, A. Archi, R. Colotti, R.M. Di Giorgi, G. Gradenigo, B. Inghirami, P. Matiello, R. Nannuci, and M. Ragona. New prospectives in information retrieval techniques: a hypertext prototype in environmental law. In *Online Management 89, Proceedings 13th International Online Information*, pages 483–494, London, United Kingdom, 1989.

[ACG89] M. Agosti, F. Crestani, and G. Gradenigo. Towards Data Modelling in Information Retrieval. *Journal of Information Science*, 15(6):307–319, 1989.

[ACG91] M. Agosti, R. Colotti, and G. Gradenigo. A two-level hypertext retrieval model for legal data. In A. Bookstein, Y. Chiaramella, G. Salton, and V.V. Raghavan, editors, *Proceedings of the 14th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 316–325, Chicago, Illinois, October 1991. ACM Press.

[AGM92] M. Agosti, G. Gradenigo, and P.G. Marchetti. A hypertext environment for interacting with large textual databases. *Information Processing & Management*, 28(3):371–387, 1992.

[Alt99] Digital's AltaVista search engine project. http://www.altavista.com/, Last verified on: 1st of February, 1999.

[AMC95] M. Agosti, M. Melucci, and F. Crestani. Automatic Authoring and Construction of Hypermedia for Information Retrieval. *ACM Multimedia Systems*, pages 15–24, 1995.

[AMM97] P. Atzeni, G. Mecca, and P. Merialdo. To Weave the Web. In *Proceedings of the Twenty-third International Conference on Very Large Data Bases*, Athens, Greece, September 1997.

[Are93] Y. Arens, et al. Retrieving and integrating data from multiple information sources. *International Journal of Cooperative Information Systems*, 2, 1993.

[BH95] R. Burke and K.J. Hammond. Combining Databases and Knowledgebases for Assisted Browsing. In *Proceedings of the AAAI'95 Symposium on Information Gathering from Distributed, Heterogeneous Environments*, Palo Alto, California, March 1995.

[BHP94] M. Bright, A. Hurson, and S. Pakzad. Automated resolution of semantic heterogeneity in multidatabases. *ACM Transactions on Database Systems*, 19(2), 1994.

[BL97] P.D. Bruza and B. van Linder. Preferential Models of Refinement Paths in Query by Navigation. IJCAI-97 Workshop on AI and Digital Libraries, 1997. Available from http://www.icis.qut.edu.au/~bruza/pubs.html.

[Bow95] C.M. Bowman, et al. Harvest: A Scalable, Customizable Discovery and Access System. Techical report cu-cs 732-94 (revised march 1995), University of Colorado - Boulder, Computer Science Department, Boulder, Colorado, March 1995.

[Bra98] T. Brasethvik. A semantic modeling approach to metadata. *Internet Research*, 8(5):377–386, 1998.

[Bru90] P.D. Bruza. Hyperindices: A Novel Aid for Searching in Hypermedia. In A. Rizk, N. Streitz, and J. Andre, editors, *Proceedings of the European Conference on Hypertext - ECHT 90*, pages 109–122, Cambridge, United Kingdom, 1990. Cambridge University Press.

[BW90] P.D. Bruza and Th.P. van der Weide. Two Level Hypermedia - An Improved Architecture for Hypertext. In A.M. Tjoa and R. Wagner, editors, *Proceedings of the Data Base and Expert System Applications Conference (DEXA 90)*, pages 76–83, Vienna, Austria, 1990. Springer-Verlag.

[BW95] F.C. Berger and Th.P. van der Weide. A Feedback Mechanism for Query by Navigation. In R. Sacks-Davis and J. Zobel, editors, *Proceedings of the Sixth Australasian Database Conference, ADC'95*, volume 17(2) of *Australian Computer Science Communications*, pages 56–65, Adelaide, Australia, January 1995.

[CDA97] S. Castano and V. De Antonellis. Semantic dictionary design for database interoperability. In *Thirteenth International Conference on Data Engineering*, pages 43–54, Birmingham, United Kingdom, April 1997.

[Cra78] T. Craven. Linked phrase indexing. *Information Processing & Management*, 14(6):469–476, 1978.

[Dub99] Dublin Core Metadata Initiative. http://purl.org/DC/, Last verified on: 1st of February, 1999.

[Eve81] B. Everitt. *Cluster Analysis*. Heinemann Educational Books, United Kingdom, 1981.

[FEFP95] R. Fikes, R. Engelmore, A. Farquhar, and W. Pratt. Network-based Information Brokers. In *Proceedings of the AAAI'95 Symposium on Information Gathering from Distributed, Heterogeneous Environments*, Palo Alto, California, March 1995.

[Gru92] T.R. Gruber. Ontolingua: A mechanism to support portable ontologies. Technical Report KSL 91-66, Department of Computer Science, Stanford University, Stanford, California, March 1992.

[Gru93] T.R. Gruber. Towards Principles for the Design of Ontologies Used for Knowledge Sharing. In N. Guarino and R. Poli, editors, *Formal Ontology in Conceptual Analysis and Knowledge Representation*, Deventer, The Netherlands, 1993. Kluwer Academic Publishers.

[HPW96] A.H.M. ter Hofstede, H.A. Proper, and Th.P. van der Weide. Query formulation as an information retrieval problem. *The Computer Journal*, 39(4):255–274, September 1996.

[KH95] D. Kuokka and L. Harad. Supporting Information Retrieval via Matchmaking. In *Proceedings of the AAAI'95 Symposium on Information Gathering from Distributed, Heterogeneous Environments*, Palo Alto, California, March 1995.

[Kir98] J. Kirriemuir, et al. Cross-searching subject gateways. *D-Lib Magazine*, January 1998.

[KM97] J. Kahng and D. McLeod. Dynamic classificational ontologies: Mediation of information sharing in cooperative federated database systems. In M. P. Papazoglou and G. Schlageter, editors, *Cooperative Information Systems: Trends and Directions*, pages 179–203, London, United Kingdom, 1997. Academic Press.

[Kow97] G. Kowalski. *Information Retrieval Systems: Theory and Implementation*. Kluwer Academic Publishers, Deventer, The Netherlands, 1997.

[KS97] V. Kashyap and A. Sheth. Semantic heterogeneity in global information systems: the role of metadata, context and ontologies. In M. P. Papazoglou and G. Schlageter, editors, *Cooperative Information Systems: Trends and Directions*, pages 139–178, London, United Kingdom, 1997. Academic Press.

[LRO96] A. Levy, A. Rajaraman, and J.J. Ordille. Querying Heterogeneous Information Sources using Source Descriptions. In *Proceedings of the Twenty-third International Conference on Very Large Data Bases*, Bombay, India, September 1996.

[MBP95] S. Milliner, A. Bouguettaya, and M. Papazoglou. A Scalable Architecture for Autonomous Heterogeneous Database Interactions. In *Proceedings of the 21th VLDB Conference*, Zurich, Switzerland, September 1995.

[ML94] M.L. Mauldin and J.R. Levitt. Web-agent related Research at the CMT. In *Proceedings of the Seventeenth Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval*, Dublin, Ireland, July 1994.

[MMW96] S. Milliner, M. Papazoglou M., and H. Weigand. Linguistic Tool based Information Elicitation in Large Heterogeneous Database Networks. In R.P. van de Riet, J.F.M. Burg, and A.J. van der Vos, editors, *Proceedings of the Second Workshop on Applications of Natural Language to Databases (NLDB'96)*, Amsterdam, The Netherlands, June 1996.

[MRF$^+$90] G.A. Miller, Beckwith R., C. Fellbaum, D. Gross, and K.J. Miller. Introduction to wordnet: An on-line lexical database. *Journal of Lexicography*, 3(4):234–244, 1990.

[MS95] D. McLeod and A. Si. The design and experimental evaluation of an information discovery mechanism for networks of autonomous database systems. In *Eleventh International Conference on Data Engineering*, pages 15–24, Taiwan, February 1995.

[Mya92] S.H. Myaeng. Using Conceptual Graphs for Information Retrieval: A Framework for Representation and Flexible Inferencing. In *Proceedings of the Symposium on Document Analysis and Information Retrieval*, pages 102–116, Las Vegas, Nevada, March 1992.

[OMN99] OMNI, Organizing Medical Networked Information. http://omni.ac.uk/, Last verified on: 1st of February, 1999.

[ONL94] T. Oates, N. Nagendra, and V. Lesser. Cooperative information gathering: A distributed problem solving approach. Technical report 94-66 (version 2), Department of Computer Science, University of Massachussets, Amherst, Massachussets, 1994.

[Pap95] M.P. Papazoglou. Unraveling the Semantics of Conceptual Schemas. *Communications of the ACM*, 38(9), September 1995.

[PB99] H.A. Proper and P.D. Bruza. What is Information Discovery About? *Journal of the American Society for Information Science*, 50(9):737–750, July 1999.

[PIC99] Platform for Internet Content Selection (PICS). http://www.w3.org/PICS/, Last verified on: 1st of February, 1999.

[Pin94] B. Pinkerton. Finding what People Want: Experiences with the WebCrawler. In *Proceedings of the 1st International Conference on the WWW*, Geneva, Switzerland, May 1994.

[PM96] M.P. Papazoglou and S. Milliner. Pro-active Information Elicitation in Wide-area Information Networks. In *Proceedings of the International Symposium on Cooperative Database Systems for Advanced Applications.* World Scientific, Japan, December 1996.

[PM98] M.P. Papazoglou and S. Milliner. Subject-based organization of the information space in multi-database networks. In *Proceedings of the Tenth International Conference CAiSE'98 on Advanced Information Systems Engineering*, Lecture Notes in Computer Science, pages 251–272, Pisa, Italy, 1998. Springer-Verlag.

[RB89] R. Rada and E. Bicknell. Ranking documents based on a thesaurus. *Journal of the American Society for Information Science*, 40(5), May 1989.

[SB88] G.E Salton and C. Buckley. Term-weighting approaches in automatic text retrieval. *Information Retrieval and Management*, 24(5):513–523, 1988.

[Sch96] R.B. Schatz et al. Interactive Term Suggestion for Users of Digital Libraries. In *First ACM International Conference on Digital Libraries*, pages 126–133, Bethesda, Maryland, March 1996.

[She95] M.A. Sheldon. *Content Routing: A Scalable Architecture for Network-Based Information Discovery.* PhD thesis, Massachussets Institute of Technology, Boston, Massachussets, December 1995.

[SOS99] SOSIG: The Social Science Information Gateway. http://www.sosig.ac.uk/, Last verified on: 1st of February, 1999.

[Sow84] J.F. Sowa. *Conceptual Structures: Information Processing in Mind and Machine.* Addison-Wesley, Reading, Massachusetts, 1984.

[WGMD95] S. Weibel, J. Godby, E. Miller, and R. Danierl. Metadata Workshop Report. Dublin, Ohio, March 1995.

[Wie96] R. Wiess, et al. HyPersuit: A Hierarchical Network search Engine that Exploits Content-link Hypertext Clustering. In *7th ACM Conference on Hypertext*, Washington D.C., March 1996.