Subject-Based Organization of the Information Space in Multi-database Networks

Michael P. Papazoglou[1] and Steven Milliner[2]

[1] Tilburg University, INFOLAB,
P.O. Box 90153, 5000 LE Tilburg,
The Netherlands
mikep@kub.n

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

Abstract. Rapid growth in the volume of network-available data, complexity, diversity and terminological fluctuations, at different data sources, render network-accessible information increasingly difficult to achieve. The situation is particularly cumbersome for users of multi-database systems who are expected to have prior detailed knowledge of the definition and uses of the information content in these systems.

This paper presents a conceptual organization of the information space across collections of component systems in multi-databases that provides serendipity, exploration and contextualization 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 data a natural expectation for a variety of users. A limiting factor however, is the difficulty in providing coherent access and correlation of data that originate from diverse widely-distributed data sources. This is an involved process not only due to the sheer volume of information available, but also because of heterogeneity in naming conventions, meanings and modes of data usage. Differences in data descriptions, abstraction levels, and precise meanings of terms being used in disparate data 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 database sites over time. Users are thus presented with the problem of gaining adequate knowledge of a potentially huge, complex 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 [24]. This expectation is quite unreasonable in large distributed systems.

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 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 and includes: (i) *Determining* the information needs of users by means of different term suggestions; (ii) *Locating* candidate database sources that address these needs; (iii) *Selecting* schema items of interest from these sources; and finally, (iv) *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 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 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. These and related issues are addressed herein.

This paper is organized as follows. Section 2 presents related work, while section 3 discusses a logical organization for the semantic cross correlation of metadata information from component databases in a multi-database system. Section 4 presents clustering techniques, while section 5 outlines navigation and querying mechanisms. Finally, section 6 presents our conclusions and future work. This work is an extension and elaboration of some early ideas outlined in [14] and [15]. In [14] 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 [15] we described IR techniques and algorithms used for the physical clustering of databases. In this paper we concentrate on the details of logical database formation, according to subject, based on a common terminology context and present navigation and querying techniques.

2 Finding Information: An Overview

In this section a number of techniques from different fields for locating information are discussed.

Web-based Resource Discovery

The use of the World Wide Web (WWW) has led to the development of a variety of search engines which attempt to locate a large number of WWW documents by indexing large portions of the Web. These tools recursively enumerate hypertext links starting with some known documents. We can classify search engines into two broad categories: *centralized index* and *content-based* search engines.

Centralized index search engines such as Lycos [11], Web Crawler [19] 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 [23] and Harvest [2]. These systems generate summarized descriptions (*content labels*) of the contents of information servers.

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 [26] by generalizing collections so that they may span documents from various servers.

The Harvest information discovery and access system [2] 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 [9]. 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) [25], 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) [16] 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, lies in the way that these perform indexing. Alta Vista indexes individual pages and not resources. For example, a large document consisting of many Web pages 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.

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 [3], [1], [12], [4]. Construction of conceptual ontologies on the basis of domain-specific terminologies and formalisms that can be mapped to description logics are also discussed in [8]. Some of the issues relating to the identification of semantically related information can be found in [3], 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 [7] 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.

3 System Organization

In order to improve efficient searching/elicitation of schema information in large multi-database networks, the first task is to partition the multi-database information space into distinct subject (domain-specific) categories meaningful to database users. Categorization and subject classification 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 [22]. Domain-specific partitioning organizes databases in logical clusters and makes searches more directed, meaningful and efficient. In addition, a subject directory created as a result of domain-specific database categorization can also provide subject-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 elicitation. Firstly, an organization of *raw* data must be introduced for the discovery of data inter-relationships. Topic classification schemes for this purpose as they summarize related information subspaces together. Secondly, this organizational structure must itself be scalable – that is: interactions with it must be scalable, and maintenance of it 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 topic-based architecture for multi-databases.

3.1 Subject-based Database Clustering

Our approach to information elicitation in large database networks relies on logically partitioning the multi-database schema information space into distinct subject (topic) categories meaningful to users. This occurs by creating logical objects called Generic Concepts (GCs) to achieve explicit semantic clustering of associated component database schema elements. Database-content clustering automatically computes sets of related component databases – via their exported meta-data terms – and associates them with an appropriate generic concept, see Figure 1. Generic concepts essentially represent centroids of the inter–component database schema information space – around which databases cluster – and are engineered to describe a particular domain (generic concepts were termed "Global Concepts" in previous work [14]).

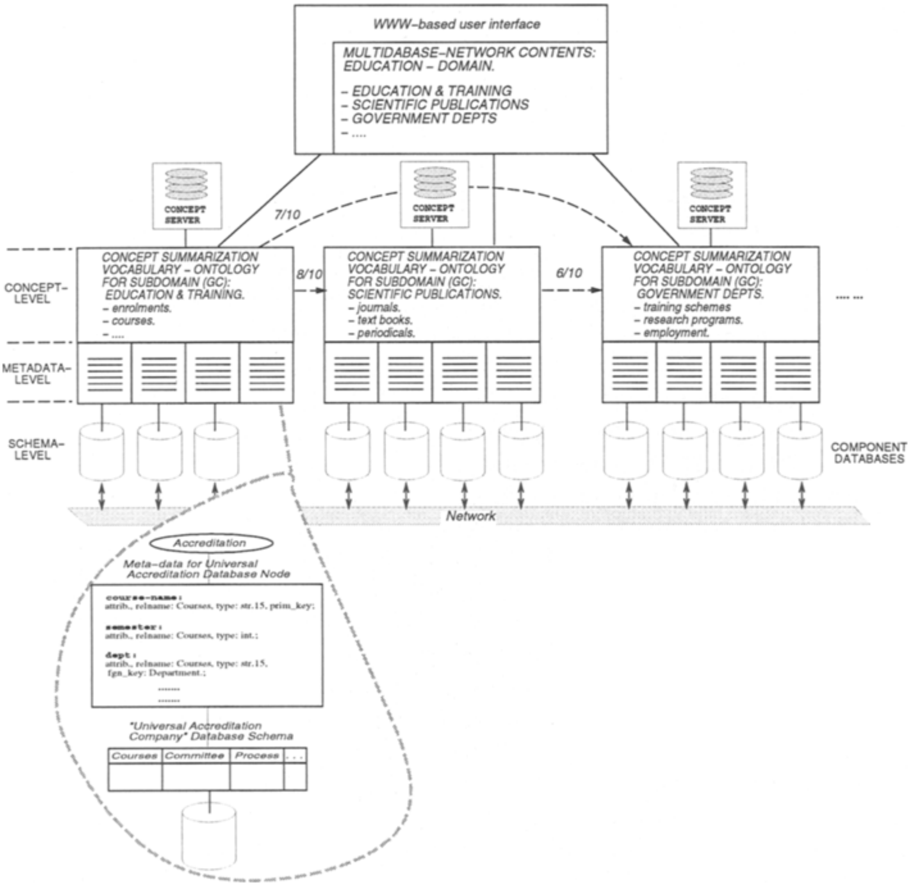

WWW-based user interface
MULTIDABASE-NETWORK CONTENTS: EDUCATION - DOMAIN.
- EDUCATION & TRAINING
- SCIENTIFIC PUBLICATIONS
- GOVERNMENT DEPTS
- ....
CONCEPT SERVER
CONCEPT SERVER
CONCEPT SERVER
7/10
8/10
6/10
CONCEPT-LEVEL
METADATA-LEVEL
SCHEMA-LEVEL
CONCEPT SUMMARIZATION VOCABULARY - ONTOLOGY FOR SUBDOMAIN (GC): EDUCATION & TRAINING.
- enrolments.
- courses.
- ....
CONCEPT SUMMARIZATION VOCABULARY - ONTOLOGY FOR SUBDOMAIN (GC): SCIENTIFIC PUBLICATIONS.
- journals.
- text books.
- periodicals.
CONCEPT SUMMARIZATION VOCABULARY - ONTOLOGY FOR SUBDOMAIN (GC): GOVERNMENT DEPTS.
- training schemes
- research programs.
- employment.
....
COMPONENT DATABASES
Network
Accreditation
Meta-data for Universal Accreditation Database Node
course-name:
attrib., relname: Courses, type: str.15, prim_key;
semester:
attrib., relname: Courses, type: int.;
dept:
attrib., relname: Courses, type: str.15, fgn_key: Department.;
....
....
"Universal Accreditation Company" Database Schema
Courses
Committee
Process
. . .

Fig. 1. Partitioning a multi-database information space into generic concepts.

To participate in GC-structured database network, a component database must export part of its *meta-data* to the other databases in the network. This means that the component database administrator must specify which part of the database meta-data can be made available for sharing with other database systems in the network. We refer to these meta-data as the *exported meta-data.* Figure 1 shows a sample database, called the Universal_Accreditation_Company database, along with a partial representation of its meta-data. Although meta-data contain also physical definitions such as definitions of views, ownership, authorization privileges, indexes and access patterns, these (except for authorization privileges) are not important for inclusion in the GC level.

A GC organized multi-database schema information space can be viewed as a Web-space that encompasses collections of exported meta-data. A GC organized multi-database schema 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. Databases in this network may connect to more than one GCs if they strongly relate to their content. To circumvent terminology fluctuations we provide a standard vocabulary for interacting with the GCs. In this way we create a *concept space* (information sub-space) for a specific topic category. The concept space constitutes a type of summarization or synoptic topic knowledge regarding a particular domain, e.g., *education and training*, *publications*, *government tertiary-related departments*, etc, and is stored in a GC, see Figure 1. This clustering mechanism results in grouping exported meta-data elements from diverse databases that share important common properties onto a generic concept, associating these properties with the GC representation, and regarding the GC as an atomic unit. A GC is thus a form of a logical object whose purpose is to cross-correlate, collate, and summarize the meta-data descriptions of semantically related network-accessible data. This scheme provides an appropriate frame of reference for both component database schema term indexing and user instigated searches. With this scheme navigation can be considered as browsing through databases exclusively at a topic-level i.e., from topic area to topic area such as from educational training, to publications, government departments and so on.

To put the organization of a concept space into perspective, we consider the case of a domain based on educational information provided by a large number of interconnected databases as shown in Figure 1. This figure also illustrates how a component database (Accreditation) – which provides information about accreditation of courses and cross-institutional subjects, various private/public educational training information and other similar or related data – is connected to the GC network. In its original form the Accreditation database, maintains information only on education service providers, their courses, accreditation committee members, accreditation processes and related information. Figure 1 shows the Accreditation database along with a partial representation of its associated meta-data and schema. It also illustrates how this component database may become part of a larger network by establishing weighted links to GCs implementing related areas of interest. Consequently, the Accreditation database is not only able to source appropriate information on its subject matter but also to provide *matching* information about enrollment programs, training schemes, government programs, research activities and publication data.

By linking to a certain GC, databases agree to associate with each other and thus inter–component database organization is achieved implicitly. In addition, GCs are interconnected by weighted links (called *content links*) to make the searches more directed and meaningful, see Figure 1. Each of the component databases may also link less strongly (e.g., 7/10) to other GCs which have their own associated cluster of database nodes. Presently, the degree of relatedness between GCs is decided by database administrators. Accordingly, a single database, e.g., Universal_Accreditation_Company, may be simultaneously involved in several clusters of databases (information sub-spaces) to varying degrees, as dictated by the weights of its content links to the various GCs. The resulting GC structure forms a massive dynamic network, resembling a cluster–based *associative network* (a variant of semantic networks that uses numerically weighted similarity links).

Overall a networked information system may be viewed in terms of three logical levels. The bottom level (Figure 1) corresponds to the schemas of the component databases. The middle level represents *exported meta–data* for the database schemas. The top most level corresponds to the *concept space* (GC) level. This level contains abstract dynamic objects which implement the clustering of related portions of the underlying component meta-data and materialize the GCs in an object-oriented form. Figure 1 illustrates that there is a one-to-one correspondence between database schemas and their meta-data representations, while an entire collection of exported meta-data corresponds to a single concept-space. This three-tier architecture is the key ingredient to information elicitation in distributed, scalable systems. It provides the ability to describe varying levels of aggregated database sources and the granularity of the information components, i.e., exported meta-data terms, that comprise them. It generates a semantic hierarchy for database schema terms in layers of increasing semantic detail (i.e., from the name of a term contained in a database schema, to its structural description in the meta-data level, and finally to the concept space level where the entire semantic context – as well as patterns of usage – of a term can be found). Searches always target the richest semantic level, viz. GC level, and percolate to the schema level in order to provide access to the contents of a component database, see section 5.

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 [17]. Although GCs 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 [3], [4]. The advantage of forming conceptual database clusters is that searches are goal–driven[3] 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.

3.2 Generic Concept Characteristics

Individual GCs are useful for browsing and searching large database collections because they organize the information space. For example, the Education and Training Providers concept space provides a common terminology basis upon which database nodes dealing with enrollments, courses, training, accreditation, etc, (see Figure 1), achieve knowledge of each others information content.

A GC is a definitional or schematic construct: it corresponds to a class hierarchy depicting all terms within the topic sampled by the GC. The GC structure is illustrated in Figure 2. This figure shows that each GC is characterized by its name and the context of its terms (term hierarchy and term descriptions) for each specific topic. Terms within a GC are shown to have a distinct meaning (sense) and context. This concept space consists of abstract descriptions of terms in the domain, term senses, relationships between these terms, composition of terms, terminology descriptions, hypernym, hyponym, antonyms-of, part-of, member-of (and the inverses), pertains-to relations, contextual usage (narrative descriptions), a list of keywords, and other domain specific information, that apply to the entire collection of members of a GC, Figure 2. Hence, the GC structure is akin to an associative thesaurus and on-line lexicon (created automatically for each topic category). Thesaurus-assisted explanations created for each subject-based abstraction (GC-based information subspace) serve as a means of disambiguating term meanings, and addressing terminology and semantic problems. Therefore, the GC assists the user to find where a specific term that the user has requested lies in its conceptual space and allows users to pick other term descriptions semantically related to the requested term.

Operations on a GC object include mapping services which map GC provided terms to semantically related terms in the component databases. They also include summarization services which summarize the exported meta-data from component databases to implement a GC. Summarization services aggregate networks of exported meta-data terms (one per component database). This mechanism is described in a later section.

An example of the GUI for some of the the terms included in the educational GC is given in Figure 3. Here, we assume that a user who searches the entries in the educational GC is interested in the term course and wishes to gain more insight into its semantic context. The first step after entering the term is to choose the *senses* from the list the GC lexicographic substrate provides. The sense number returned is then associated with the term (as is the case with all other words in the term description). For example, Figure 3 shows that the

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

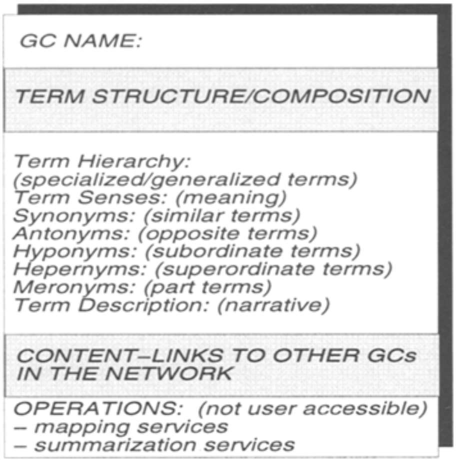

GC NAME:
TERM STRUCTURE/COMPOSITION
Term Hierarchy:
(specialized/generalized terms)
Term Senses: (meaning)
Synonyms: (similar terms)
Antonyms: (opposite terms)
Hyponyms: (subordinate terms)
Hepernyms: (superordinate terms)
Meronyms: (part terms)
Term Description: (narrative)
CONTENT–LINKS TO OTHER GCs
IN THE NETWORK
OPERATIONS: (not user accessible)
– mapping services
– summarization services

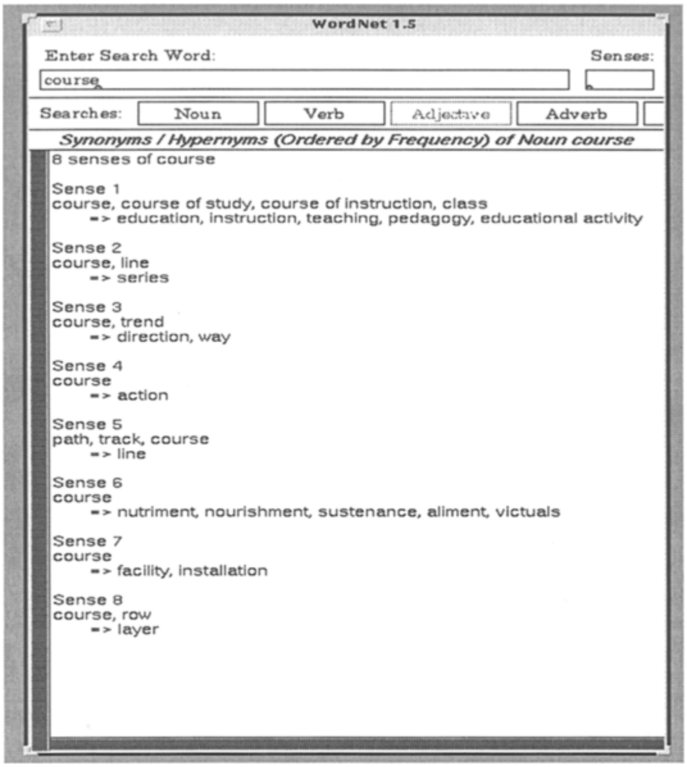
Fig. 2. Generic concept structure.

term course has eight senses (meanings), but once the domain of discourse is limited to study (education), then only one of the eight can occur. Figure 4 which is an expansion of the specific term chosen, shows how the GC provides the necessary information needed for the contextual representation, i.e., meaning, of a specific term. Other factors such as the context of usage (not shown here due to space limitations) can be combined with its contextual representation to restrict the search space. Thus the user gets a complete picture regarding the semantic context of this and associated terms (see Figure 4) and is free to pick up a desired term(s) which would eventually lead him/her to candidate component data sources. Term entries in this GUI are mapped by means of the mapping services of a GC to the relevant schema terms found in component databases (in the same GC).

Information contained in the GCs is stored in an information-repository that resides at a *concept server* associated with and accessible by the databases clustered around a specific conceptual information space (GC), see Figure 1. The concept server implements an individual GC, performing abstraction and summarization operations on its underlying meta-data. This information-repository contains thus a rich domain model that enables describing properties of the database sources clustered around a GC.

4 Representation and Clustering of Schema Meta-data

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



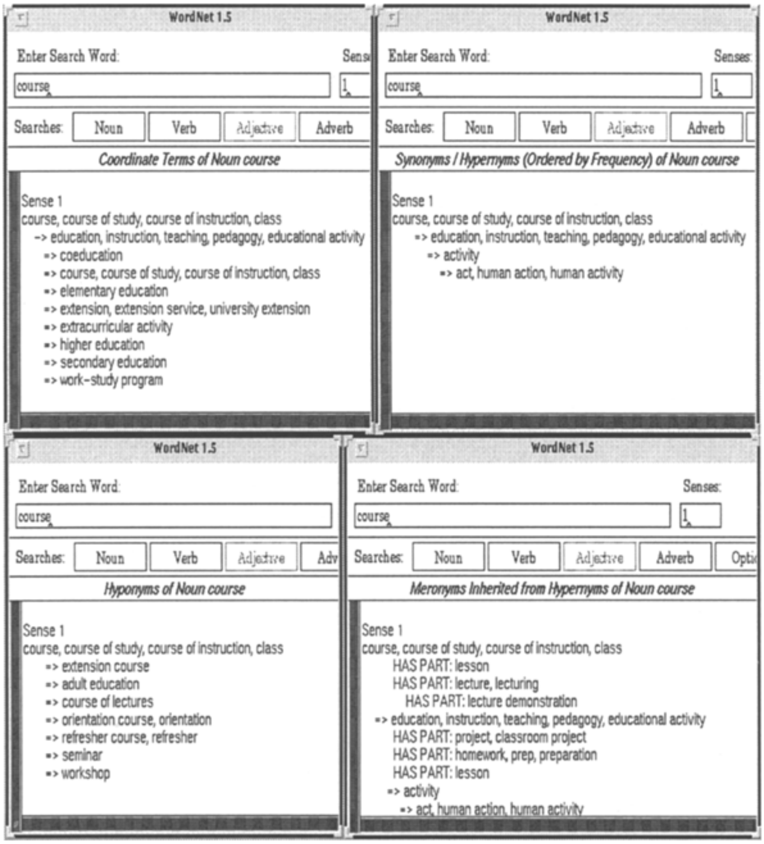
Synonyms / Hypernyms (Ordered by Frequency) of Noun course

```
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
```

Fig. 3. Choosing the meaning of the term course.

4.1 Describing the Meta-Data Content of a Database Node

In order to initialy cluster component databases a high level description of the meta-data content of a database must first be developed. To demonstrate this consider the previous example of the Universal_Accreditation database, which deals with academic institutions and accreditation processes. This database contains entities such as courses, committees, (accreditation) processes, etc. We use a variant of an information retrieval (IR) technique called, *star technique*, where a term is selected and then all terms related to it are placed in a class[10]. Terms not yet in a class are selected as new seeds until all terms are assigned to a class. The variant of the star technique that we are using starts with a term represented as an abstract class (*term descriptor class*), then an additional term that is related to the term selected is represented as a another class and is connected to the selected term. The new term is then selected as a pivot and the process is repeated until no new terms can be added. In this way a *context graph* created for a specific database schema. For example, the context graph for the Universal_Accreditation component database (Figure 5) contains nodes



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
```

Fig. 4. More contextual information regarding the term course.

which correspond to the abstract term descriptor classes committee, institutions, courses etc., while the context graph edges depict inter–connections (association, generalization, specialization or containment) between the terms within this particular database. Term interrelations are determined on the basis of a reference lexicographic substrate that underlies all the GCs in the network. For this purpose we use the lexicographic system[4] WordNet [13] that supports semantic term matching through the use of an extensive network of word meanings of terms connected by a variety of textual and semantic relations.

To facilitate clustering and discovery of information, we require that a component database (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 database; associations between meta-data terms in the form of a semantic-

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

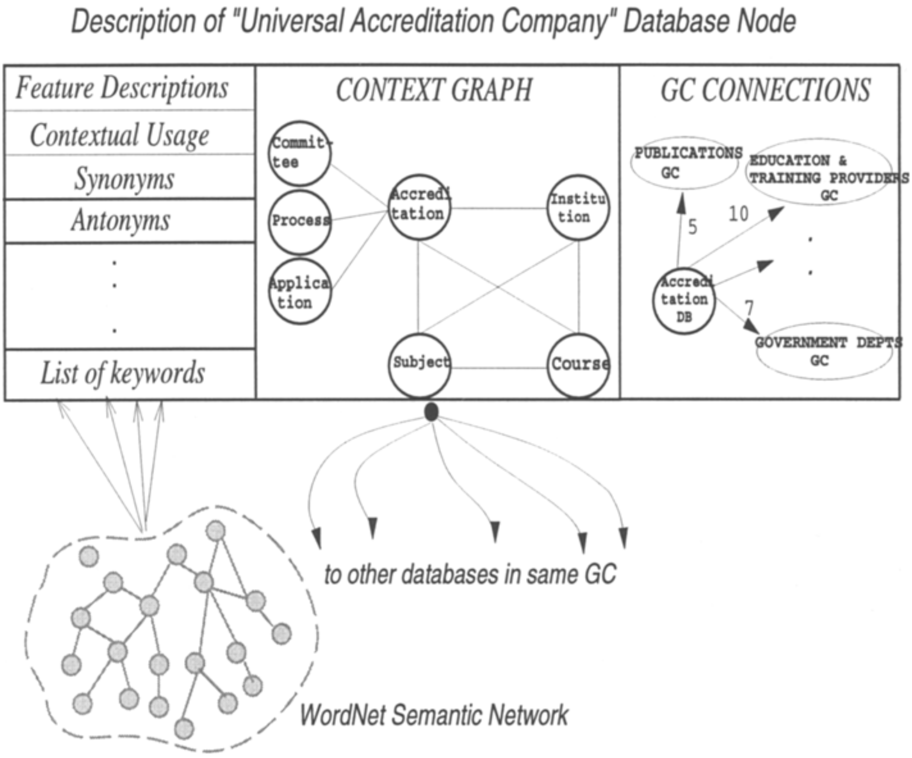


Fig. 5. Describing a component database.

net; and finally, links from these descriptions to other related databases 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 5 illustrates that each database node contains the following sections: a *feature descriptions*, a *context graph*, and a *GC connections* section. The feature descriptions section contains information about terms, composition of terms, remarks about the meaning of terms, hypernym, hyponym, antonyms-of, part-of, member-of (and the inverses), pertains-to relations and lists of keywords. This section may also include certain details such as: geographical location, access authorization and usage roles, explanations regarding corporate term usage and definitions, domains of applicability 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, 3 and 4. The context graph section contains a non–directed graph which connects term synopses (in the form of *term descriptor classes*) found in the Universal_Accreditation database schema. Except for viewing purposes, the term descriptor nodes and their link structure are used in the clustering of databases to form the generic concepts. Each of the term descriptor nodes defines (in conjunction with its respective entry in the feature descriptions window) a common structured vocabulary of terms – describing the term in question, e.g., course, – and a specification of term relationships within that particular subject. Finally, the GC connection section shows how the Universal_Accreditation database is related, i.e., content link weights, to other GCs in the network.

4.2 Similarity-based Clustering of Database Nodes

Similarity-based clustering of database schemas organizes databases into related groups based on the terms (term descriptor nodes) 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 term 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 term/link similarity-based algorithm because it is relatively easy to implement and avoids generating very large clusters.

Term-based Similarity: this is calculated using cluster analysis techniques [5] to identify co–occurrence probabilities – representing the degree of similarity – between two discrete terms. Our similarity metric is based on the meaning of the collection of terms representing the topical context (viz. semantic-levels) of a particular term, e.g., course, and the synonyms of these, see Figure 3. The comparison is based on: a conversion of each context graph node (e.g., term descriptor) Committee, Process, Subject, Course, etc. (see Figure 5) to a corresponding matrix of noun terms (containing the entire topical context of a term); and a subsequent comparison of terms within these matrixes.

A matrix $a_{n,m}$ of (noun) terms, representing the topical context of a particular term, $a_{i,1}$ (course say), will correspond to the name of the term descriptor in the context graph. The synonyms of this term will be $a_{i,2}$, $a_{i,3}$ ... $a_{i,m}$ (course-of-study, course-of-lectures). Terms $a_{i-x,j}$ $(x > 0)$, e.g., education, educational-activity, will be more general than terms $a_{i,j}$, while terms $a_{i+x,j}$ will be more specific, e.g., CS-course. In the final step, all synonyms for these terms are generated to produce the node's a complete *topical description matrix* $a_{n,m}$ for a specific term.

Similarity analysis is mainly based on statistical co–occurrences of term descriptor objects based on techniques which has been successfully used for automatic thesaurus generation of textual databases [5], [21]. In fact we base our term-based similarity on the improved *cosine* formula [21] 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 [13].

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 [20]. 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 (term descriptors) 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 [15].

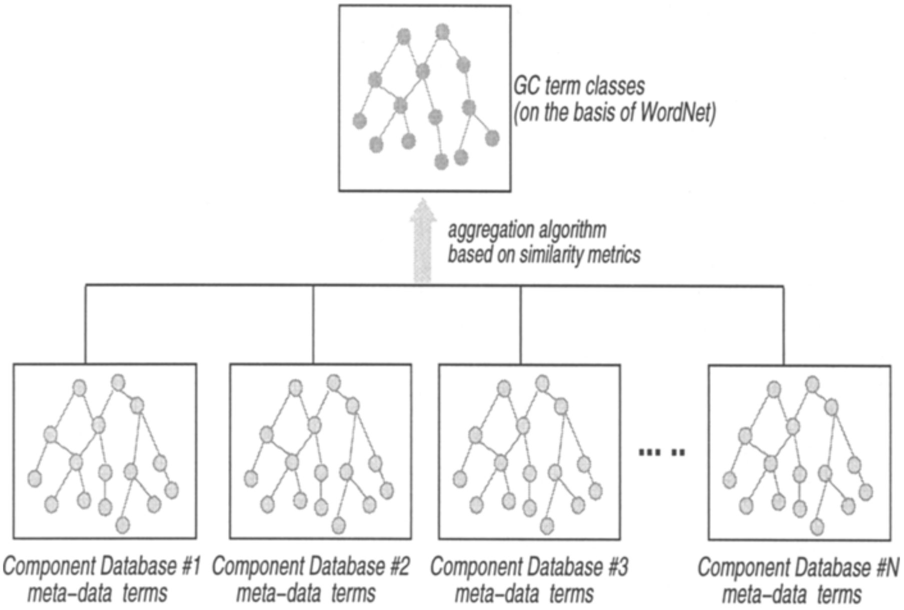


Fig. 6. Clustering interrelated component schema terms.

Once similarity between nodes has been established context graphs are aggregated to create GCs. The aggregation of the context graphs from various component databases, results in the clustering of inter–related database schemas, see

Figure 6. The aggregation algorithm employed does not integrate the aggregates, as is the usual case with other approaches [8], but rather links descriptor classes at the GC level with corresponding term descriptor classes in its underlying cluster of database context graphs. Again this association is performed on the basis of the reference lexicographic substrate (WorNet). For each database cluster, a GC is created to represent the area of interest (or concept) that the group embodies, e.g., Education and Training Providers GC for the Employee Training, Accreditation, and Government Education Center databases as depicted in Figure 2.

5 Schema Term 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.

5.1 Navigation Techniques

There are two basic modes in which searching of the system may be organized. These search modes depend upon the nature of the information a user is attempting to access, and how this information relates to the database that user is operating from. Serendipity, exploration and contextualization are supported by means of indexing based upon terms contained in the component database context graphs. 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 weights of content links of this database to other GCs in the network. We refer to this form of exploration as *concept-driven.* Concept-driven querying is the subject of a previous publication [18]. In this paper we will concentrate on index-driven exploration and on the querying of schema-related information.

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 and is related to the dynamic indexing schemes and incremental discovery of information requirements for information elicitation. In order to traverse the index a user will have to decide on a number of key request terms, and then select synonyms or more general (and perhaps more specific) derivatives of these key terms. The resulting query structure - generated on the basis of terms extracted from WordNet entries - can then be compared against the context graph structure of component databases.

User specified term comparison starts at the top of the GC generated index and gradually percolates down to the required level of specificity by following the terms at each level. Figure 7 depicts this process in terms of a user query

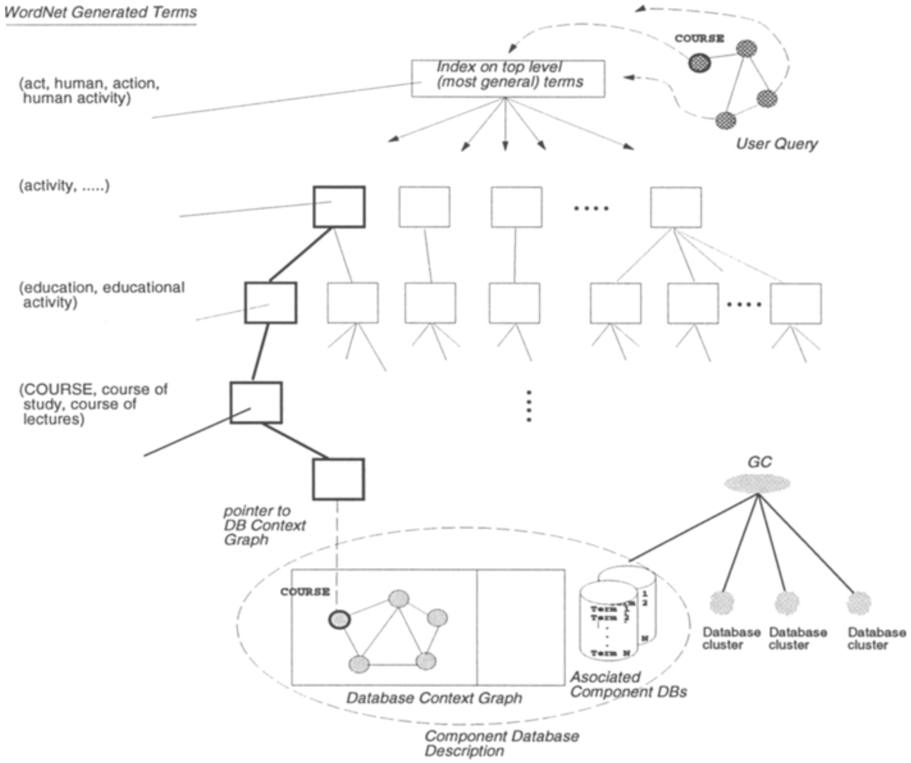

WordNet Generated Terms
(act, human, action, human activity)
Index on top level (most general) terms
COURSE
User Query
(activity, ...)
(education, educational activity)
(COURSE, course of study, course of lectures)
pointer to DB Context Graph
GC
COURSE
Term 1
Term 2
Term N
Database cluster
Database cluster
Database cluster
Database Context Graph
Asociated Component DBs
Component Database Description

Fig. 7. Accessing the index.

requesting information about courses at various institutions. Here we assume that the user has already specified that s/he is interested in the contents of the Education & Training GC. The graph of the user's query supplied terms contains a node Course and this term is used to traverse the GC generated index and arrive at the collection of databases which include this term (or its aliases) in their own descriptions. The index-driven navigation process starts with the most general terms possible, e.g., act, human activity, that correspond to the requested query term (course). These terms are generated by the GC (via the WordNet) and are presented to the user for selection. Once the user has selected a general term, most specific terms are revealed, e.g., education. Once a GC term matching a user supplied term is selected, a link is established with the context graphs of all component databases containing the desired term (or its aliases). In this way the user can obtain contextual information and possibly a partial view of potentially matching databases and then s/he can decide whether a candidate database is useful or not. This hierarchical form of schema term navigation guarantees that a user supplied term correlates semantically with the content of the component databases underlying a GC cluster. The process is then repeated for all the other terms in the user's query graph (i.e. the remaining unlabeled nodes in Figure 7). Thus, by matching the user query graph nodes to semantically equivalent GC terms, we can infer a number of component databases that are most closely associated to the user query.

5.2 Querying of Domain Meta-Data

When the user needs to further explore the search target, *intensional*, or schema queries [17] – 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 component databases (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 GC in Figure 4 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 component database 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 100 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 component databases 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 wȁs 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 databases 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.

This last experiment is important as it must be remembered that different people may be responsible for the construction of different database descriptions. Thus, the automatic clustering must be relatively robust in terms of the way different people may describe the same object. It is expected that, given all descriptions will be generated using the same thesaurus, the system should prove relatively good at detecting differing descriptions of a single object.

Currently, experiments have been performed using a "full" database description involving the synonyms, generalizations and terms senses, as well as the structural relationships between these terms, see Figure 4. Initialy, the term matching component was based on the standard similarity metric proposed by Dice [5], and the structural similarity was based on the notion of spreading activation energy [15]. 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

This paper described the fundamental aspects of a scalable, semantically oriented, configurable distributed information infrastructure that supports information discovery and retrieval across subject domains in networked databases. The proposed logical architecture extracts semantics from database schemas and creates dynamic clusters of databases centered around common topics interest (viz the generic concepts). 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. To surmount semantic-drifts, the terminology problem and enhance database retrieval, alternative search terms and term senses are suggested to users. This architecture enables users to gather and rearrange information from multiple networked databases in an intuitive and easily understandable manner. 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 GC 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 [5]. 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 GCs may need to be updated (strengthened or weakened) over time depending on interaction, new GCs may be formed, and existing GCs may need to merge. The next suite of experiments to be performed will deal with the characteristics of the link weight update and GC split/merge processes. From this policies will be developed (e.g. delayed/batch updating of GC information), and then evaluated.

References

1. Arens Y., et al. "Retrieving and Integrating Data from Multiple Information Sources", *Int'l Journal of Cooperative Information Systems*, **2**, 2, (1993).

2. Bowman. C. M., et al. "Harvest: A Scalable, Customizable Discovery and Access System", Univ. of Colorado - Boulder, CS Dept., techn. report CU-CS 732-94, (1995).
3. Bright M., Hurson A., Pakzad S. "Automated Resolution of Semantic Heterogeneity in Multidatabases" ACM ToDS, **19**, 2, (1994).
4. Castano S., De Antonellis V. "Semantic Dictionary Design for Database Interoperability", *13th Int'l Conf. on Data Engineering*, Birmingham, April (1997), 43–54.
5. Everitt B. "Cluster Analysis", *Heinemann Educational Books Ltd.*, Great Britain, (1981).
6. Kahle B., Medlar A. "An Information System for Corporate Users: Wide Area Information Servers", The Interoperability Report, **5**, 111, (1991).
7. Kahng J., McLeod D. "Dynamic Classificational Ontologies: Mediation of Information Sharing in Cooperative Federated Database Systems", in *Cooperative Information Systems: Trends and Directions*, Papazoglou M. P., Schlageter G. (eds), Academic-Press (1997) 179–203.
8. Kashyap V., Sheth A. "Semantic Heterogeneity in Global Information Systems: the Role of Metadata, Context and Ontologies", in *Cooperative Information Systems: Trends and Directions*, Papazoglou M. P., Schlageter G. (eds), Academic-Press (1997) 139–178.
9. Kirriemuir J. et al., "Cross-Searching Subject Gateways", *D-Lib Magazine*, January (1998).
10. Kowalski G. "Information Retrieval Systems: Theory and Implementation", *Kluwer Academic Publishers*, (1997).
11. Mauldin L.M., Levitt J.R. "Web-agent related Research at the CMT", *Procs. ACM Special Interest Group on Networked Information Discovery and retrieval: SIGIR'94*, August (1994).
12. McLeod D., Si A. "The Design and Experimental Evaluation of an Information Discovery Mechanism for Networks of Autonomous Database Systems", *11th Int'l Conf. on Data Engineering*, Taiwan, Feb. (1995) 15–24.
13. Miller G. "WordNet: A Lexical Database for English", Communications of ACM, **38**, 11, Nov. (1995).
14. Milliner S., Bouguettaya A., Papazoglou M.P. "A Scalable Architecture for Autonomous Heterogeneous Database Interactions", *21 Int'l Conference on Very Large Databases*, Zurich, Switzerland, Sept. (1995).
15. Milliner S., Papazoglou M., Weigand H. "Linguistic Tool based Information Elicitation in Large Heterogeneous Database Networks", *NLDB '96 Natural Language and Databases Workshop*, Amsterdam, June (1996).
16. "OMNI, Organizing Medical Networked Information", http://omni.ac.uk/
17. Papazoglou M.P. "Unraveling the Semantics of Conceptual Schemas", Communications of ACM, **38**, 9, Sept. (1995).
18. Papazoglou M.P., Milliner S. "Pro-active Information Elicitation in Wide-area Information Networks", *Procs. of the Int'l Symposium on Cooperative Database Systems for Advanced Applications*, World Scientific, Japan, Dec. (1996).
19. Pinkerton B. "Finding what People Want: Experiences with the WebCrawler", Procs. 1st Int'l Conference on the WWW, Geneva, May (1994).
20. Rada R., Bicknell E. "Ranking Documents Based on a Thesaurus", Journal of the American Society for Information Science, **40**, 5, May (1989).
21. Salton G.E, Buckley C. "Term-Weighting Approaches in Automatic Text Retrieval", *Information Retrieval and Management*, **24**, 5, (1988), 513–523.

22. Schatz R.B., et. al "Interactive Term Suggestion for Users of Digital Libraries", *1st ACM International Conf. on Digital Libraries*, Bethesda MD, March (1996), 126–133.
23. Sheldon M.A. "Content Routing: A Scalable Architecture for Network-Based Information Discovery", PhD thesis, MIT, Dec. (1995).
24. Sheth A., Larson P. "Federated Database Systems for Managing Distributed, Heterogeneous and Autonomous Databases". *Computing Surveys*, **22**, 3, Sept (1990).
25. "SOSIG: The Social Science Information Gateway", http://www.sosig.ac.uk/
26. Wiess R., et al. "HyPersuit: A Hierarchical Network search Engine that Exploits Content-link Hypertext Clustering", *7th ACM Conf. on Hypertext*, Washington DC., March (1996).