DEEP FOCUS

HYDRA-HEADED METADATA

After Sept. 11, authorities said information-stove-piping by intelligence agencies was one of the biggest stumbling blocks in the fight against terrorism. Now, two leading researchers discuss different approaches to merging government files, and cracking open their secrets.

By: Jamie Callan, Language Technologies Institute, Carnegie Mellon University

W. Bruce Croft, Computer Science Department, University of Massachusetts, Amherst

Eduard Hovy, Digital Government Research Center, Information Sciences Institute, University of Southern California

Introduction

The terrorist events of September 11 reminded everyone of the need for accurate and timely government intelligence. Some of the information that might prevent disasters is secret, and therefore inaccessible. But in many cases, the information is present somewhere, and freely available. The problem is getting hold of it in a usable form.

Unfortunately, while present-day government in almost all its branches has collected, analyzed and stored information, most of it non-uniform. Information is all over the place, in hundreds of different formats and systems and versions. You don't know where to find it, how to access it, or how to convert it to a format you can work with once you actually have it.

One of the principal problems facing those trying to standardize non-homogeneous data sets is variation in terminology. For example, what one agency calls *salary*, another might call *income*, and a third calls *wages*, while using *salary* to mean something else entirely. Example: one agency might calculate monthly average prices of unleaded gasoline in California by measuring wholesale rates each month, while another measures prices at selected pumps weekly and averages them. The results will differ, but both will be called "*average monthly gasoline prices in California"*.

Clearly, this state of affairs causes confusion for not only Government workers, but also for journalists, congressional staffers, students, the general public, and intelligence officers. All would benefit from government information systems that locate, retrieve, and integrate desired information quickly, handling transparently the details of which databases contain the information or in what format it is presented. No system should expect its patrons to trust its results unquestioningly, so these information systems should also make it easy to examine the relationships among documents and/or databases with similar conten,t if desired.

The basis for any new system is *metadata*, that is data that describes data or collections of data. The Dewey Decimal system, the Library of Congress Subject Headings, Medical Subject Headings (MESH), and many other controlled vocabularies (sometimes called *ontologies*) are all familiar forms of metadata. Each document is catalogued by a small number of terms from the controlled vocabulary, as is each information request, and matching them is very simple.

But practical experience has shown that integrating vast and disparate term sets and data definitions to create new forms of metadata is fraught with difficulty. The U.S. Government has funded several metadata initiatives, including the Government Information Locator Service (GILS) and the Advanced Search Facility (ASF) (http://www.gils.net/, http://asf.gils.net/). These projects perform exemplary work in establishing a structure of cooperation and standards between agencies, including structural information (formats, encodings, links). However, they do not focus on the actual creation of metadata, nor do they define the algorithms needed to generate it.

Experience with traditional forms of metadata has shown that it is expensive and time-consuming to produce, that people (e.g., authors) often resist creating it when there is no immediate or direct benefit, and that information-seekers often find it difficult to relate their requests to pre-specified ontologies or controlled vocabularies. Generating a common ontology for a domain also tends to be controversial. New standards for communicating metadata, such as XML, do nothing to address the underlying issue of where it originates. Controlled vocabularies and relatively static ontologies are not solid foundations for information systems that must cover a wide range of subjects, support rapid integration of new information, be easy for the general population to use, and can only be maintained at moderate expense. Large-scale use of metadata requires new answers to fundamental questions.

Recently, the Digital Government program of the National Science Foundation has funded a number of projects to address the challenge of integrating large, heterogeneous, widely distributed and disparate Government data collections. In this paper, we describe two complementary approaches: large ontology-based data access planning using small domain models semi-automatically acquired, and dynamic metadata creation from language models.

Ontology-based data access planning

The DGRC[1] Energy Data Collection (EDC) Project was started in the National Science Foundation's Digital Government program in 1999. The EDC project is working with representatives of Federal and State statistics agencies and other organizations to build a system for disseminating statistical data from the Census Bureau, the Bureau of Labor Statistics (BLS), the Energy Information Administration (EIA) of the Department of Energy (DoE), and the California Energy Commission (CEC). An example of the kind of information the project is working with appears at the EIA's site http://www.eia.doe.gov. The EIA's extensive monthly energy data is open to the public, and the site receives hundreds of thousands of hits a month, even though only the last few years of data are available, and only as downloads of standard web (HTML) pages or as prepared PDF documents. The current facility thus supports only limited access to a potentially rich data source.

The EDC Project is developing effective methods to identify and describe the contents of databases so that useful information can be accurately and efficiently located by everyone, even those with no expertise in the domain.

In order to standardize terminology and provide single-point access, the project has extended a USC/ISI taxonomy to incorporate new energy-related terms, organized in small domain-related taxonomies called *domain models*. Instead of building new ontologies for each domain or

[1] The Digital Government Research Center (DGRC; www.dgrc.org) was established to perform Information Technology research as needed in Government. The DGRC consists of faculty, staff, and students at the Information Science Institute (ISI) of the University of Southern California and Columbia University's Computer Science Department and its Center for Research on Information Access.

database, this project takes an existing, large-scale and fairly neutral ontology SENSUS (Knight and Luk, 1994) and extends it with just enough information to model the contents of the new domain or database. This incremental method makes domain modeling by humans much easier and at the same time allows one to locate cross-domain inconsistencies and terminology clashes.

SENSUS, built at USC/ISI, contains approximately 70,000 terms linked together into a subsumption (*is-a*) network, with additional links for part-of, pertains-to, and so on. SENSUS is a rearrangement and extension of WordNet (Fellbaum, 1998), built at Princeton University on general cognitive principles, then re-taxonomized under the Penman Upper Model (Bateman et al., 1989), which was constructed at ISI to support natural language processing. Most of SENSUS' content is identical to WordNet 1.5. SENSUS can be accessed using the ontology browsers DINO at http://edc.isi.edu:8011/dino and Ontosaurus at http://mozart.isi.edu:8003/sensus/sensus_frame.html (Swartout et al., 1996).

Dynamic metadata creation using text mining

In another NSF digital government project, researchers at the University of Massachusetts at Amherst and Carnegie Mellon University are pursuing a new approach in which metadata is automatically generated, based on *language models* instead of ontologies or controlled vocabularies. Simple language models represent basic vocabulary and frequency information; more complex language models represent phrases, names, and other speech patterns, as used in the texts surrounding the data collections (data descriptions, glossaries, technical publications, etc.).

Language models can form a more detailed representation of document or database contents than a few controlled vocabulary terms could be expected to achieve. Language models also enable a system to generate descriptions (metadata) directly from the content of its databases, without trying to match database contents to a controlled vocabulary. Language models are easily updated as information is added to a database, they support an unlimited range of subjects (because they are generated directly from database contents), and they enable a wide range of information-seeking activities.

Techniques for automatically generating database descriptions (Gravano, 1997; Callan, 1995) have been primarily aimed at the problem of *resource location*. This is the problem of determining, given a user's query, which databases are the most appropriate or the most likely to contain relevant data. Resource location, also called *collection selection* in the research literature, is the basis of much recent research on *distributed retrieval* (Gravano, 1994; Callan, 1995; Voorhees, 1995; Xu, 1998). In the U-Mass/ Carnegie Mellon research, it is assumed that after databases have been selected, local search engines will evaluate the query and the local search results will be merged for presentation to the user.

The most common technique for collection selection is to represent a collection as a word histogram, which is usually a list of words that occur in the collection and their associated frequencies. The virtual document representations described in Callan, Lu and Croft (1995) and Xu and Callan (1998) are examples. The word histograms are indexed and the resulting data structure is called the collection selection index (Callan, Lu and Croft, 1995). Collection selection consists of simply ranking the word histograms against a query in the same way as ranking ordinary documents. Most other techniques are very similar. Such a technique is a simple modification of document retrieval. Documents and collections of documents are, however, very

different and techniques that work well for one problem may not work well for the other. A document usually deals with only one topic but a typical collection can deal with many topics.

Matching the words in a query with different topics in a collection can cause failure in collection selection. Suppose a collection contains documents about fruits and computers. Matching the query "Apple Computer"' against the histogram of the collection will produce a high similarity even though none of the documents are relevant. Heterogeneous collections therefore make the simple technique of matching a query against word histograms less effective for collection selection. In fact, if all collections are sufficiently heterogeneous, matching a query consisting of a few common words with word histograms can even produce random collection selection because statistically all histograms are almost identical with respect to the query. Xu and Callan (1998) showed that ineffective collection selection could seriously degrade the performance of distributed retrieval.

Ontologies as metadata: definition and acquisition

To retrieve information dispersed among multiple sources, users need familiarity with their contents and structure, query languages and location. A person (or system) with need for distributed information must ultimately break down a retrieval task into a collection of specific queries to databases and other sources of information (e.g., analysis programs). With a large number of sources, individuals typically do not possess the knowledge or time required to determine how to find and process the information they need. Even if they did, performing the necessary tasks would be time-consuming and prone to error.

The DGRC project's approach to integrating statistical databases builds on research performed by the SIMS group at ISI (Arens et al., 1996). SIMS assumes that the system designer specifies a global model of the application domain and describes the contents of each source (database, web server, etc.) in terms of this global model. SIMS software provides a single point of access for all the information: the user expresses queries without needing to know anything about the individual sources. SIMS translates the user's high-level request, expressed in a subset of SQL, into a *query plan* (Ambite and Knoblock, 2000), a series of operations including queries to sources of relevant data and manipulations of the data. Queries are expressed internally in the Loom knowledge representation language (MacGregor, 1990).

Since its start in 1999, the project has incorporated over 50,000 tables, from sources in various formats, (including Oracle and Microsoft Access databases, HTML web forms and pages, and PDF files), collected from the Energy Information Administration, the Census Bureau, the Bureau of Labor Statistics, and the California Energy Commission. A large amount of the information is in the form of semi-structured web pages. These web sources were 'wrapped' automatically using technology from the Ariadne system (Muslea et al., 1998). Ariadne allows a developer to mark up example web pages using a demonstration-based GUI. Then the system inductively learns a landmark grammar that is used to extract the marked-up fields from similar pages and generates all the necessary wrapper code. The resulting wrapper acts as a simple relational database that accepts parametrically-defined SQL and dynamically retrieves data from the associated web pages and forms.

In SIMS, each of these data sources, whether natively relational or wrapped by Ariadne, is modeled by associating it with an appropriate domain-level concept description. A set of approximately 500 domain terms, organized in 10 sub-hierarchies, constitutes the domain model required so far for the EDC domain. A fragment of the EDC domain model is shown in Figure 1. This model describes time-series data about different gasoline products. A time series is defined by a set of dimensions such as *product type* (e.g., unleaded gasoline, premium gasoline), *property* measured (e.g., price, volume), *area* of the measure (e.g., USA, California), *unit* of measure, etc.

Each of the time series in the sources is described by using specific values for each of the hierarchical dimensions. For example, a particular source may be described as providing the monthly prices (based on the consumer price index) of premium unleaded gasoline for the state of California. The dimensions can be seen as metadata that describes the series. The actual data is modeled as a set of measurements (i.e., date and value pairs). The domain model also describes whether a source has footnotes for some of the data. The answer to a query will also return the footnote data associated with the corresponding tuples if so requested.

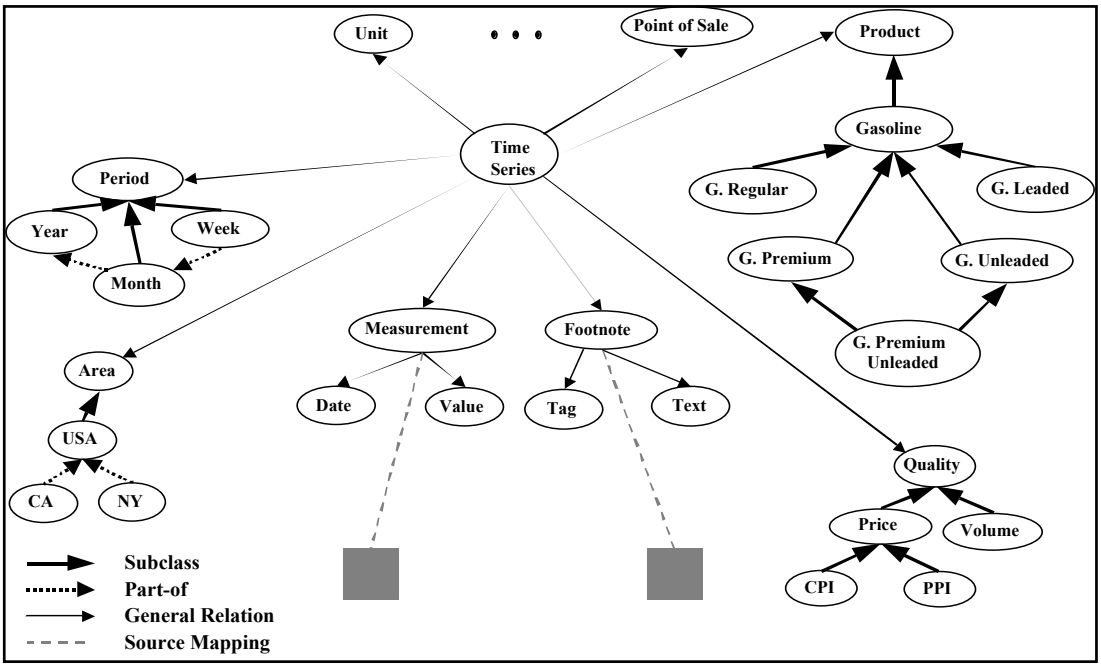

Figure 1. Fragment of the EDC Gasoline Domain Model

In order to achieve cross-domain coverage and to identify incompatibilities across domains, small domain models such as this are linked into SENSUS. The project is developing semi-automated methods of producing small domain-specific models for new domains and databases, and for automatically relating these domain ontologies into the overarching large one. These methods include algorithms to extract domain terms from online glossaries and domain texts (Klavans et al., 2000), to cluster them, and to automatically align them to SENSUS terms (Hovy et al., 2001).

In particular, the EDC models have been linked into SENSUS as follows. Each of the retrievable time series, along with each of the ten dimensional values, has been added to SENSUS as an ontological concept in its own right; the relationships between series and dimensional values have been reified as SENSUS relations as well (e.g., has-product-type, area-of, etc.). Much of our research was devoted to performing this linking semi-automatically (see below). Using tools that facilitate the construction of wrappers and the semi-automatic description of sources is critical to scale mediator systems to the very large number of information sources that are available from government agencies in a cost-effective fashion.

The ontology for the EDC project has the structure shown in Figure 2, using two types of links. The first, called *generally-associated-with*, holds between concepts in the ontology and domain model concepts, allowing the user while browsing to rapidly proceed from high-level concepts to the concepts associated with real data in the databases. This is a loose association, intended to link domain concepts into many potentially relevant general concepts. The second is a strict logical mapping from the domain concept to the appropriate database models, intended to support reasoning about logical equivalence of alternative data items.

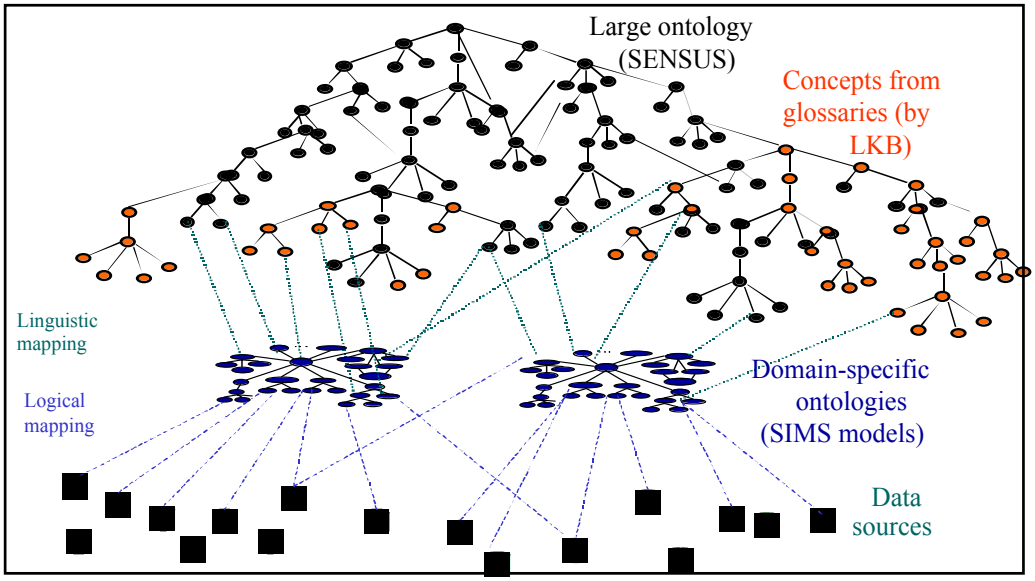

Figure 2. Ontology and Domain Models

Language models as metadata: definition and acquisition

The phrase "language model" is used by the speech recognition community to refer to a probability distribution that captures the statistical regularities of the generation of language (Yamron, 1997). Generally speaking, language models for speech attempt to predict the probability of the next word in an ordered sequence. For the purposes of document retrieval, one can model occurrences at the document level without regard to sequential effects and obtain good retrieval results. The approach to retrieval described in Ponte and Croft (1998) is to infer a language model for each document and to estimate the probability of generating the query according to each of these models. Documents are then ranked according to these probabilities. Most retrieval systems use term frequency, document frequency and document length statistics. Typically these are used to compute a *tf.idf* weight that is used in indexing (Robertson and Walker, 1994). In the language modeling approach, collection statistics such as term frequency, document length and document frequency are integral parts of the language model and do not have to be included in an ad hoc manner. In the digital government project, the researchers are studying the use of language models to represent information resources rather than individual documents.

Database selection algorithms depend upon accurate language models, but the problem of acquiring accurate language models has received little attention. The state-of-the-art is the proposed STARTS extension to Z39.50 (Gravano, 1997), which requires each database provider to provide language models upon request. STARTS is a *cooperative protocol*, because success depends upon each database being *able* to cooperate, choosing to cooperate, representing its

contents accurately. Cooperative protocols are unable to deal with old ("legacy") databases, databases with no incentive to cooperate, and databases that misrepresent their contents, either accidentally or intentionally.

Cooperative protocols are also based on the assumption that compatible language models are provided for different databases, but this assumption is rarely true in practice. Information Retrieval systems use many types of lexical processing, such as stopword removal, stemming, case folding, acronym recognition, and specialized indexing for proper names, to name a few, that make language models created by different providers incompatible. If two databases each report 1,000 occurrences of the stem 'apple', it is impossible to know which contains more documents about Apple computers. Word frequency statistics cannot be compared without knowing how they were derived, but it is impractical to document every assumption and decision that went into the production of each word frequency statistic, even if a site were willing to do so.

The weaknesses of cooperative protocols make them unsuitable for environments in which database control is divided between many parties. Another solution is required.

The UMass-CMU research team is developing a new solution, called *query-based sampling*. Query-based sampling assumes only that each available database is capable of accepting simple queries and returning a relatively small number of matching documents. A sequence of queries, each returning a biased sample of the database, is constructed that, collectively, provide a relatively unbiased sample of the database. Language models are then constructed from the set of sampled documents. The initial research shows that the resulting "learned" language models are similar to the actual language models ("perfect information") (Callan, 1998a), and that the resulting language models enable relatively accurate database selection (Callan, 1998b).

Once the data sources have been represented by a set of language models, the obvious next issue to be addressed is how these language models will actually be used.

Automatic Selection of Structured Databases

Much of the government information available in the today's networked environment consist of text documents, such as the Congressional Record and laws in the THOMAS legislative access system, or the many HTML pages that contain descriptions of government agencies, policies, regulations, and procedures. The distributed retrieval techniques discussed in the previous section have also been designed to handle large numbers of databases, which may contain few or many text documents. If this approach to generating metadata for cross-database search is to be extended to cover a significant proportion of government information, however, techniques for dealing with structured information must be developed. Structured information, particularly in the form of tables and relational databases, is a critical part of the government information infrastructure. There are, for example, about 70 government agencies that generate statistical information and 10 whose primary function is to generate and analyze statistics. Given a query such as "exporting computers to China", a cross-database search should return not only documents on export policies and procedures, but also tables showing the volume of this type of export over the last few years.

Pyreddy and Croft (1997) developed a retrieval approach for tabular data that has been extracted from documents. Tables are indexed as text documents, with extra weight assigned to words from captions, row and column headings. A retrieval experiment showed that, using this representation, an IR system was able to effectively retrieve tables in response to queries. In the current proposal, we are dealing with databases that contain many tables, or with tables that

contain large amounts of data (as in a relational table), but the research of Pyreddy and Croft indicates that a language model approach to representing these databases should be investigated. In the case of a database that contains a large number of tables, such as those generated by the Department of Commerce, clustering techniques similar to those used for text databases could generate language models. The table representation used for clustering would be based on the words occurring in captions, headings, and the body of the table. The research issues that need to be addressed involve understanding which estimation techniques are the most effective for creating metadata for table databases.

U-Mass and Carnegie Mellon researchers believe that it should be possible to extend the capability of the centralized component of the distributed environment (the collection selection server) to also support linkage and browsing as well as research. Figure 3 gives an overview of a possible architecture. The databases shown are examples of sources for government-related information.

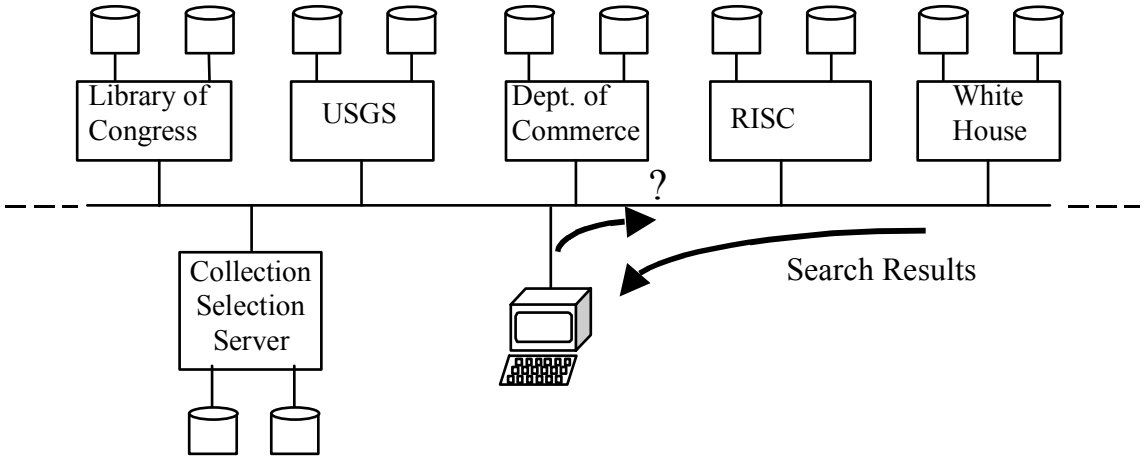

Figure 3. Architecture for Cross-Database Search and Linkage

By providing a linkage and browsing facility in the collection selection server, the person doing the searching will be able to see what types of information are available and what alternative sources exist before any search is done in the local databases. Given that very general searches in such environments can retrieve large amounts of information, being able to see potential groupings of information will give the searcher the opportunity to refine their query and focus the search. Another possible mode of operation would be to carry out a distributed search and retrieve the top-ranked items. When the searcher sees an item that is specifically relevant to their interests, they will be able to ask to see the "related items" from this and other databases. This is an extension of the "more documents like this" feature found in a number of web search services. The effectiveness of either mode will depend on a number of factors, including how language models can be summarized and presented to searchers. The primary issue, however, is the degree to which groups of information objects can be successfully linked based on language model representations.

Conclusion

Both these projects represent significant advances in research on database access. If successful, their solutions to different aspects of the problem may possibly one day be merged, enabling the database integrator to produce language models, and from them to induce metadata schemas that support database access planning and seamless data integration and delivery from non-homogeneous, and potentially very different, data sources. Even if the projects are only partially successful, the outcome should benefit not only to the Government, but any organization with many disparate data sources.

References

J. Allan, J. Callan, M. Sanderson, J. Xu, S. Wegmann. "INQUERY and TREC-7" In *the Proceedings of the 7th TREC conference (TREC-7)* published by NIST, 1998.

Ballesteros, L. and Croft, W.B. Resolving Ambiguity for Cross-Language Retrieval, *Proceedings of the 21st ACM SIGIR*, 64-71, 1998.

A. Bookstein, D. Swanson, 1976. "Probabilistic models for automatic indexing." *Journal of the American Society for Information Science*, 25(5), p. 312-318.

J. Callan, M. Connell, A. Du. "Query-based sampling of text databases". Technical Report IR-154, Center for Intelligent Information Retrieval, University of Massachusetts. 1999.

J. Callan, M. Connell, A. Du. "Automatic discovery of language models for text databases." In *Proceedings of the 1999 ACM-SIGMOD International Conference on Management of Data*, pp. 479-490. 1999. ACM.

J.P. Callan, Z. Lu, and W.B. Croft. "Searching distributed collections with inference networks." In *Proceedings of the Eighteenth Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp 21-28, 1995. ACM.

Hsinchun Chen, Andrea Houston, Robin R. Sewell, Bruce R. Schatz: Internet Browsing and Searching: User Evaluations of Category Map and Concept Space Techniques. JASIS 49(7): 582-603 (1998)

D.W. Embley, Y. Jiang, and Y.K. Ng. "Record-boundary discovery in Web documents." In *Proceedings of the 1999 ACM-SIGMOD International Conference on Management of Data*, pp. 467-479. 1999. ACM.

J.C. French, A.L. Powell, J. Callan, C.L. Viles, T. Emmitt, and K.J. Prey. "Comparing the performance of database selection algorithms." To appear in *the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval.* 1999.

J.C. French, J.C. Powell, C.L. Viles, T. Emmitt, and K.J. Prey. "Evaluating database selection techniques: A testbed and experiment." In *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM, 1998.

J. Goldstein, M. Kantrowitz, and J. Carbonell. "Summarizing Text Documents: Sentence Selection and Evaluation Metrics." To appear in *the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval.* 1999.

L. Gravano, K. Change, H. García-Molina, and A. Paepcke. "STARTS Stanford proposal for Internet meta-searching". In *Proceedings of the ACM-SIGMOD International Conference on Management of Data*, 1997.

L. Gravano and H. García-Molina. "Generalizing GLOSS to vector-space databases and broker hierarchies." In *Proceedings of the 21st International Conference on Very Large Databases (VLDB)*, pp. 78-89, 1995.

L. Gravano, H. García-Molina, and A. Tomasic. "The effectiveness of GLOSS for the text database discovery problem." In *Proceedings of SIGMOD 94*, pp 126--137. ACM, 1994.

S.P. Harter, 1975. " A probabilistic approach to automatic keyword indexing." *Journal of the American Society for Information Science*, 24.

H.S. Heaps. *Information Retrieval: Computational and Theoretical Aspects.* Academic Press: New York. 1978.

Hearst, M. and Pedersen, J., Re-examining the Cluster Hypothesis: Scatter/Gather on Retrieval Results, *Proceedings of the 19th ACM SIGIR*, 76-84, 1996.

Larkey, Leah "A Patent Search and Classification System," to appear in the *Proceedings of Digital Libraries (DL 99)*, Berkeley, CA, Aug. 11-14, (1999).

D. Lewis and P. Hayes (editors), Special issue on Text Categorization, *ACM Transactions on Information Systems*, 12(3), (1994).

Leouski, J. Allan, 1997. "Evaluating a visual navigation system for a digital library," *Proceedings of the Second European Conference on Research and Technology for Digital Libraries*.

C. Manning, H. Schutze, 1999. *Foundations of Statistical Natural Language Processing*. MIT Press.

J. Ponte, W.B. Croft, 1998. "A Language Modeling Approach to Information Retrieval." Proceedings of the 21st International Conference on Research and Development in Information Retrieval, p. 275-281.

J. Ponte, 1998. *A Language Modeling Approach to Information Retrieval.* Ph.D. thesis, Computer Science Department, University of Massachusetts.

P. Pyreddy and W.B. Croft, "TINTIN: A System for Retrieval in Text Tables", *Proceedings of the ACM Conference on Digital Libraries*, 193-200, (1997).

Van Rijsbergen, C.J. *Information Retrieval*, Second edition, Butterworths, London, 1979.

S.E. Robertson, K. Sparck Jones, 1977. "Relevance weighting of search terms." *Journal of the American Society of Information Science*, 27.

S.E. Robertson, S. Walker, 1994. "Some simple effect approximations to the 2-Poisson model for probabilistic weighted retrieval." *Proceedings of ACM SIGIR '94*, p. 232-241.

B.W. Silverman, 1985. *Density Estimation for Statistics and Data Analysis*. John Wiley and Sons.

A. Tombros, M. Sanderson, 1998. "Advantages of query-biased summaries in information retrieval." *Proceedings of ACM SIGIR '98*, p. 2-10.

H.R. Turtle, W.B. Croft, 1992. "A comparison of text retrieval models." *Computer Journal*, 35(3), p. 279-290.

E.M. Voorhees, N.K. Gupta, and B. Johnson-Laird. "Learning collection fusion strategies." In *Proceedings of the Eighteenth Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp 172-179, Seattle, 1995. ACM.

J. Xu and W.B. Croft. "Query Expansion Using Local and Global Document Analysis," in *Proceedings of the 19th International Conference on Research and Development in Information Retrieval (SIGIR 96)* , Zurich, Switzerland, pp. 4-11.

J. Xu and J. Callan. "Effective retrieval of distributed collections." In *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 112-120, Melbourne, 1998. ACM.

J. Xu and W.B. Croft. "Cluster-based Language Models For Distributed Retrieval." To Appear in *Proceedings of the 22nd International ACM SIGIR Conference on Research and Development in Information Retrieval*. Also available as Technical Report IR-153, Center for Intelligent Information Retrieval, University of Massachusetts. 1999.

J. Yamron, 1997. "Topic detection and tracking segmentation task." *Proceedings of the DARPA Topic Detection and Track Workshop.*

G. K. Zipf. *Human Behavior and the Principle of Least Effort: An Introduction to Human Ecology.* Addison-Wesley: Reading, MA. 1949.

Ageno, A., I. Castellon, F. Ribas, G. Rigau, H. Rodriguez, A. Samiotou. 1994. TGE: Tlink Generation Environment. *Proceedings of the 15th COLING Conference*. Kyoto, Japan.

Arens, Y., C.A. Knoblock and C.-N. Hsu. 1996. Query Processing in the SIMS Information Mediator. In A. Tate (ed), *Advanced Planning Technology*. Menlo Park: AAAI Press.

Ambite J.L. and C.A. Knoblock. 2000. Flexible and Scalable Cost-Based Query Planning in Mediators: A Transformational Approach. *Artificial Intelligence Journal*, 118 (1-2).

Bateman, J.A., Kasper, R.T., Moore, J.D., and Whitney, R.A. 1989. A General Organization of Knowledge for Natural Language Processing: The Penman Upper Model. Unpublished research report, USC/Information Sciences Institute, Marina del Rey, CA.

Fellbaum, C. 1998. (ed.) WordNet: An On-Line Lexical Database and Some of its Applications. Cambridge: MIT Press.

Harinarayan, V., A. Rajaraman, and J. D. Ullman, 1996. Implementing Data Cubes Efficiently, *Proceedings of the1996 ACM SIGMOD Conference*.

Hovy, E.H. 1998. Combining and Standardizing Large-Scale, Practical Ontologies for Machine Translation and Other Uses. *Proceedings of the 1st International Conference on Language Resources and Evaluation* (LREC). Granada, Spain.

Hovy, E.H., A. Philpot, J.-L. Ambite, and U. Ramachandran. 2000. Automating the Placement of Database Concepts into a Large Ontology. In preparation.

Klavans, J. L. 1988. COMPLEX: A Computational Lexicon for Natural Language Processing. *Proceedings of Twelfth International Conference on Computational Linguistics (COLING).* Budapest, Hungary.

Klavans, J. L., C. Jacquemin and E. Tzoukermann. 1997. "A Natural language approach to multi-word term conflation". *Proceedings of the DELOS conference* from the European Research Consortium on Information Management (ERCIM). Zurich, Switzerland.

Klavans, J. L. and Muresan S. 2000 (in press). "DEFINDER: Rule-Based Methods for the Extraction of Medical Terminology and their Associated Definitions from On-line Text". *Proceedings of 2000 American Medical Informatics Association (AMIA) Annual Symposium*, Los Angeles, California.

Knight, K. and S.K. Luk. 1994. Building a Large-Scale Knowledge Base for Machine Translation. *Proceedings of the AAAI Conference*.

MacGregor, R. 1990. The Evolving Technology of Classification-Based Knowledge Representation Systems. In John Sowa (ed.), *Principles of Semantic Networks: Explorations in the Representation of Knowledge*. Morgan Kaufmann.

Muslea, I. and S. Minton and C. A. Knoblock. 1998. Wrapper Induction for Semistructured Web-based Information Sources. Proceedings of the Conference on Automated Learning and Discovery. Pittsburgh, PA.

Okumura, A. and E.H. Hovy. 1994. Ontology Concept Association using a Bilingual Dictionary. *Proceedings of the 1st AMTA Conference*. Columbia, MD.

Rigau, G. and E. Agirre. 1995. Disambiguating Bilingual Nominal Entries against WordNet. *Proceedings of the 7th ESSLI Symposium*. Barcelona, Spain.

Swartout, W.R., R. Patil, K. Knight, and T. Russ. 1996. Toward Distributed Use of Large-Scale Ontologies. *Proceedings of the 10th Knowledge Acquisition for Knowledge-Based Systems Workshop*. Banff, Canada.