Early User—System Interaction for Database Selection in Massive Domain-Specific Online Environments

JACK G. CONRAD
Thomson Legal & Regulatory
and
JOANNE R. S. CLAUSSEN
West Group

The continued growth of very large data environments such as Westlaw and Dialog, in addition to the World Wide Web, increases the importance of effective and efficient database selection and searching. Current research focuses largely on completely autonomous and automatic selection, searching, and results merging in distributed environments. This fully automatic approach has significant deficiencies, including reliance upon thresholds below which databases with relevant documents are not searched (compromised recall). It also merges documents, often from disparate data sources that users may have discarded before their source selection task proceeded (diluted precision). We examine the impact that early user interaction can have on the process of database selection. After analyzing thousands of real user queries, we show that precision can be significantly increased when queries are categorized by the users themselves, then handled effectively by the system. Such query categorization strategies may eliminate limitations of fully automated query processing approaches. Our system harnesses the WIN search engine, a sibling to INQUERY, run against one or more authority sources when search is required. We compare our approach to one that does not recognize or utilize distinct features associated with user queries. We show that by avoiding a one-size-fits-all approach that restricts the role users can play in information discovery, database selection effectiveness can be appreciably improved.

Categories and Subject Descriptors: H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval—*search process*; *selection process*; H.5.2 [**Information Interfaces and Presentation**]: User Interfaces—*user-centered design*; *evaluation / methodology*; H.4.2 [**Information Systems Applications**]: Types of Systems—*decision support*; H.3.4 [**Information Storage and Retrieval**]: Library Automation—*large text archives*

General Terms: Design, Model, Experimentation, Performance, Evaluation, Validation

Additional Key Words and Phrases: Database selection, metadata for retrieval, structuring information to aid search and navigation, user interaction

Authors' address: J. G. Conrad, Thomson Legal & Regulatory, Research & Development, 610 Opperman Drive, St. Paul, Minnesota 55123 USA; email: {Jack.Conrad,Joanne.Claussen}@WestGroup.com.

1. INTRODUCTION

We have developed a model for improved database selection that offers the user a key role in the discovery process. The model is based on the recognition that queries can vary extensively and that techniques that treat all queries the same are bound to compromise overall performance. The experiments and evaluation described in this article focus exclusively on the resulting research system. A production implementation based upon our research and user acceptance of the production system are discussed later in this work. The foundation of our system is the WIN search engine[1] [Thompson et al. 1995; Turtle 1991, 1994], a close relative to the INQUERY engine developed at the Center for Intelligent Information Retrieval at the University of Massachusetts [Allan et al. 1997; Broglio et al. 1993]. The performance of our system has led us to question some of the underlying assumptions behind what are currently viewed as state-of-the-art database selection techniques.[2] Many of these techniques require extensive knowledge of the term and concept distribution in available collections either directly or through preliminary query-based sampling [Powell et al. 2000; Xu and Callan 1998; Callan et al. 1995; Callan and Connell 2001]. Some of these techniques suggest that a reorganization of large amounts of data, either by clustering or by topical organization, may improve overall retrieval performance [Xu and Croft 1999; Larkey et al. 2000]. In massive online data environments where the stream of incoming data or the requirements for updates can be daunting, such techniques may be rendered inapplicable because of the additional computational resources they require. Current research is also inclined to assume that a user's initial query, which may be a source selection query, also represents the user's final information request. We have found that this is not always the case.

Researchers have variously described this field as source selection, database selection, and collection selection, as well as server selection, depending on their focus. Source selection tends to remain quite broad, often with a bias towards publication source, whereas collection selection is more specific (as in a collection of textual documents). In our case, use of "database" can be misleading as it is not uncommon for a document from one of our original physical databases to be a member of two or more collections. Thus the assumption that the textual materials our collections contain are mutually exclusive does not hold. In the environment in which we operate, there may exist a CONTAINED-IN relationship between a specific collection and a larger more comprehensive collection, for instance, a database on health and medical case law for a particular state versus one on health and medical case law for all states. To remain reasonably coherent in this article, and aligned with the central thrust of this body of work, we use database selection and collection selection interchangeably as our primary research descriptors.

For effective database selection, it may be safe to assume that for general users optimization is warranted and that searching only the top-ranked collections is adequate in order to retrieve the largest sets of relevant documents. Yet

[1] WIN stands for Westlaw Is Natural.

[2] In this article, we use collection to refer to a database of textual documents.

in environments where the user-base tends to be composed of professionals who require "on point" documents in response to their queries, such compromises in recall may be unacceptable. Assumptions about data accessibility, costs to index and search collection representations, and document merging may be warranted for certain applications, yet for users from specific domains such as law or medicine, the associated shortcuts may be problematic.

In the majority of user sessions, legal researchers are searching for information from a known familiar source. As the practice of law has evolved over recent years, however, researchers are increasingly turning to extralegal sources to supplement their legal research. Information vendors such as West Group and Lexis-Nexis have supplied this demand with more business, medical, and scientific information. Yet as these information domains move away from the traditional domain of the legal researcher, information providers need to offer additional assistance in choosing the appropriate sources. Moreover, in domains where users are specialized professionals and therefore routinely more selective about their search results, such as in law or medicine, the low precision and recall sometimes associated with large-scale searches on the Web are generally unacceptable.

In mid-2000, analysis showed that there were in excess of two billion unique, publicly accessible "pages" on the Web, with an average of between 10 and 15 KB per page [Murray and Moore 2000; Heydon and Najork 1999]. With a rate of growth of over seven million new pages added per day, the Web was on track to double by mid-2001 [Murray and Moore 2000]. These figures indicate that in 2001 there were in the range of 40 to 60 terabytes of indexable text on the Web. West Group's alliance with Dialog puts their combined repositories at over 20 terabytes of data, corresponding to tens of thousands of databases. The majority of these new databases come from news and nonlegal domains, in contrast with West's historically legal focus. Although computational resources permit comprehensive searches against global indexes—thus in principle allowing users to be the final filter—the scope of the problem exacts a nontrivial cost. Recent experiments have focused on hundreds of collections, yet production environments provide over ten thousand collections, at times with an order of a million documents in each. Given that professional users generally demand more control of their search results and at the same time submit queries with drastically varying granularity, it may make sense to include user—system interaction earlier in the search process than during the final evaluation of returned search results.

To facilitate the information discovery process, we are developing a set of database selection tools. Some of them rely on collection metadata; others depend on language models based on the collections and document components [Conrad and Dabney 2001]. This toolkit approach is consistent with our view that one-size-fits-all methods will ultimately be ineffective for many types of queries. With hundreds of thousands of professional users requiring online access to tens of thousands of collections, it makes sense to examine the management of user database selection needs in a way that treats easily categorizable queries in a straightforward, less computationally expensive manner. In this article, we describe a database selection tool that leverages collection representations composed largely of metadata to address this selection problem.[3]

Another significant aspect of the model involves the contribution of users. The retrieval community has repeatedly called for an increased role for users in IR systems that are more effective than either computer-centric or user-centric approaches alone [Saracevic 1997; Fidel and Crandall 1998]. User-centric groups as a whole have increased their focus on personalized and customized presentation of information access options[4] [Kramer et al. 2000; Belkin 2000]. Recent developments in distributed IR, however, appear to have involved the user only in the formulation of the original query. To improve the performance of the search, our approach invites user collaboration in query formulation *and query categorization*. The underlying assumption of the model is that legal researchers will be quite capable of categorizing their information need into one of 8 to 10 high-level classes of queries. Users have subsequently found this approach extremely useful.

The remainder of this article is organized as follows. Section 2 reviews related work in database selection and contrasts our work with the core focus of such research. Section 3 describes our experimental methodology, including validation procedures. It also describes the substantial analysis of real user queries that forms the foundation of all subsequent investigation. Section 4 briefly addresses our collection ranking algorithms and how they are distinguished from related approaches. Section 5 discusses our experiments and how we evaluated our approach in comparison with existing methods. Section 6 examines this technique in the context of complete user information-seeking sessions. Our conclusions and description of future work are presented in Sections 7 and 8. Appendix A contains the instructions for the participants in our query category determination task. Appendix B contains a set of key legal research and practice areas leveraged in this work. Lastly, in Appendix C, we present a real user—system interactive database selection session, with associated prototype screens.

2. PREVIOUS WORK

Given the work of Callan, Gravano, French, and others, aspects of distributed search have been divided into as many as six core activities: collection identification and/or representation, query translation, collection ranking, collection selection, searching the chosen collections, and merging the results into a uniform set. In some cases, some of these activities may be reasonably clear-cut (e.g., natural language query processing); in others, they are not (e.g., collection representation). Their approaches to these issues have made considerable performance gains in terms of autonomous systems with no user interaction [Callan et al. 1995; Gravano et al. 1994; French et al. 1999; Powell et al. 2000]. These experiments leverage a considerable amount from fully automated

[3] In another work, we describe an approach that relies on production-caliber collection-based language models [Conrad et al. 2002].

[4] We take personalization to mean those added features based on information users have provided *implicitly*, and customization to mean those features based on information users have provided *explicitly* [Stellin 2000].

approaches, those that include database selection as well as document retrieval and merging. By contrast, our work more closely resembles that of Hawking and Thistlewaite [1999] as we are optimizing the selection of distributed collections (servers in their case). Yet the majority of these works also acknowledge an untapped role for user interaction in the selection process.

Gauch et al. discuss a training-based method to automatically map queries to query categories in a metasearch engine assignment context [Gauch et al. 1996; Fan and Gauch 1999]. This method employs a multiple agent architecture to categorize and broker user queries to a variety of remote search engines. The ProFusion system also tries to learn and continuously update confidence factors for improved result sets. The authors select a high-level taxonomy similar to *Yahoo!* with 14 categories. The experiments they conduct are limited and thus permit only qualitative interpretations of system performance.

Using an approach that represents a hybrid between automated techniques and user participation models, Wang et al. [2000] have proposed using a two-tiered *Yahoo!*-like concept hierarchy into which databases would be assigned based on the similarity of terms automatically generated from the *Yahoo!* hierarchy and those in a database's centroid. A potential problem with this approach, however, is that each database assigned to a lower-level category is automatically assigned to the top-level category, thus permitting a merging of specific and general category assignments. Wu et al. [2001] have more recently proposed a MetaSearch Engine that they claim is scalable in terms of both computations and storage. Their technique characterizes database terms primarily by max_tf statistics and only retains these statistics for enough databases to satisfy users' preferences (e.g., 20). The system is tuned to very short queries so as to be able to deemphasize the role of idf statistics. This approach would be more reliable were collections on the Web more static, but as collection sizes and vocabularies frequently change, the robustness of this technique and its optimizations remain an open research topic.

Other approaches have asked users to provide metadata concepts or applied thesauri with semantic links to a query, either before or after examining highly ranked source documents [Chakravarthy and Haase 1995; Dolan et al. 1996; Hearst 1994]. Park [2000] examined user—system interaction and database selection in the TREC environment, investigating whether users prefer and perform better when interacting with different databases separately with a common interface or interacting with the databases as if they were one. Her findings suggest that (1) more user control is important in a distributed environment, (2) distinct database characterization is important in supporting user choice for integration, (3) some users prefer database selection control together with merged results, and (4) the assumption that common (merged) interaction is best may be worth revisiting. Some of Park's findings actually support a number of our related discoveries, especially those involving user preference for greater control in database selection and interaction.

We have observed a number of problems when applying existing techniques in a very large-scale production environment. These techniques regularly index databases in some global form, beyond that of the individual collections. A number of experiments have shown the utility of using an indexed histogram of the terms in each collection. Another deficiency of related experiments is exemplified by the research on the TREC3 data, where queries averaged nearly 35 words (including the longer concept field) [Powell et al. 2000; Larkey et al. 2000]. For both proprietary data environments and the Web, queries of such length are rare and are therefore unrepresentative. Some of the problems we have encountered include effectively handling very short queries, optimizing large-scale searches to both determine best collections and best documents, and efficiently scaling and updating our representations to reflect actual production environment conditions.

3. EXPERIMENTAL METHODOLOGY

Our study has four phases. The *first phase* consists of the analysis and validation of legal research categories and the categorization of several thousand real user queries (Section 3.1). The *second phase* involves the exploration and development of effective means to deliver information resources for each category of query, by harnessing either search or directory navigation (Section 3.2 ff.). In the *third phase*, we enlist two sets of 450 user queries that meet certain query category criteria and run those queries against metadata authority resources (databases) derived from the previous phase (Section 5.1).[5] This phase also includes two validation steps involving real user queries, domain expert input, and correlation measures to test the reliability of the model's underlying assumptions. In the *fourth phase*, we evaluate results using completely new test query sets and compare the category-based technique with a baseline one-profile-per-collection approach (Section 5.2).

3.1 User Query Analysis

Approximately two weeks of real users' database selection descriptions were inspected. Users submitted them to a system by selecting a button labeled "Search for a Database." The queries totaled more than 8000 and represented over 7000 anonymous users. Approximately 7500 of these queries used natural language (the existing system's default); the remainder were Boolean queries that included proximity operators and field or date restrictors. The percentage of queries extracted from our query logs that somehow represent a duplication of a prior query is negligible. We found that the type of queries submitted tended to cluster around roughly 12 distinct categories (Table I; see Figure 6 for examples). These designations represent important metalevel categories.[6]

These categories include:

—document identifiers (e.g., by title or citation),

—named entities (e.g., person names or company names),

[5] We use the term *metadata authority resource* to refer to data sets developed around a specific type of query category (e.g., Courts & Government Agencies). The intent of these data sets is to effectively aid in mapping a user-categorized query to the collections most relevant to the user's information need. They are discussed more thoroughly in Sections 3.3 and 3.4.

[6] We make no claim, however, to have identified or validated any subcategories falling under these high-level designations.

Table I. Database Selection Queries by Frequency[a]

No.	Category	Distribution (%)
1.	Source or Publication (√)	48.2
2.	Legal Issue (√)	13.2
3.	Court or Gov't Agency (√)	7.5
4.	Practice or Research Area (√)	7.3
5.	Document by Citation (*)	4.5
6.	Company Name (*)	4.5
7.	Document by Title (*)	4.2
8.	Definition (√)	3.6
9.	Person Name (*)	3.0
10.	Geographic Name (√)	2.8
11.	News or Events (√)	1.8
12.	Financial Information	0.8
13.	None of the Above	1.4
14.	Category Indeterminable	1.9
	In Multiple Categories	(4.7)
	Total	100

[a] √ indicates use in final model (Section 3.2) and * indicates treatment by a parallel model.

- sources (e.g., publications or publishers),
- government entities (e.g., courts or agencies),
- legal practice or research areas (e.g., bankruptcy, estate planning, intellectual property),
- geographic (e.g., locations or regions),
- definitions (e.g., of terms or phrases),
- news (e.g., current events), and
- financial (e.g., stock market performance information).

These categories derive from common legal or business research tasks and the types of documents users commonly wish to retrieve. The length of the users' descriptions was generally too short to gain meaningful assistance from commonly used classification schemes, such as the Dewey Decimal classification. Legal users, like users in general, often bypass such schemes when retrieving legal or business information; instead, they search based on the source of the primary legal materials (cases, statutes, regulations). Various proprietary classification systems can be used, but are unlikely to provide assistance with queries as short and general as those in our sample. Any of these classification schemes would require appreciable granularity and offer too many possible assignments to assist users entering such queries.

3.1.1 *Query Category Determination.* In order to identify a comprehensive, reliable, and useful set of query categories to employ, we performed a preliminary query category determination experiment with the assistance of three legal domain experts. Each of the experts possessed a law degree as well as considerable experience working with user information requests either as a reference attorney[7] or as a query log analyst. In this exercise, we provided the three with a diverse set of 200 database selection queries. They came from a larger set of real user requests randomly selected from a query log associated with an existing database selection application. The queries were diverse in both length and specificity. The instructions given to the participants for determining categories for the user queries can be found in Appendix A. The participants were asked to supplement the sample set of user queries with their own knowledge of the domain and of associated information requests. The results from this preliminary query category determination task are shown in Table II.[8]

From the findings presented in Table II, we see that domain expert #1 suggests the least fine-grained categories and domain expert #3 contributes the most fine-grained, with domain expert #2's contributions representing something in between. Furthermore, expert #1 leaves out some areas that experts #2 and #3 address, and #3 goes into specific illustrations of some of expert #2's categories (e.g., *federal congressional materials*). Because the objective of this exercise was to determine a reasonable and useful number of complete categories, we observe that expert #1's offerings are subsumed by expert #2's whereas expert #3's are sometimes instantiations or examples of expert #2's. Given these observations, we rely on the contributions of domain expert #2 as our primary source of categories, while ensuring that any of the specific information types suggested by expert #3 would be satisfactorily covered by a slightly less fine-grained focus.

3.1.2 *Additional Observations on Database Selection Queries.*

Our study of the 8000 user queries also reveals that the variation in both query granularity and degree of abstraction is substantial. Some queries are very fine-grained and concrete (e.g., "Los Angeles City Ordinances"); others are generic and abstract (e.g., "Intellectual Property Rights"). The study demonstrates that nearly 50% of our users' queries tended to mention a source, for example, *Federal District Court Cases*, or a publication, for example, *The New York Times*. For other more generic queries it is nearly impossible to know what the user has in mind, such as when the user enters a query representing a general legal practice area or geographic location or region such as "Criminal Law" or "Alabama." For these types of queries, it might help if the user could be brought to some sort of subdirectory of information relating to such general topics. The second most frequent category, also the most abstract, is generally one of the most difficult to treat: legal issues, for example, "Right to compensation for employee slipping on wet floor?" The remainder can be characterized as being more concentrated (e.g., on person or company names), which can be handled effectively using other means. Tools for finding references and links to person names, company names, and document citations have been broadly developed [Conrad and Utt 1994; Dozier and Haschart 2000; Borlase 1999]. In this user query analysis, a number of the

[7] A reference attorney is a lawyer who has passed a bar exam in at least one of the 50 states and who answers online legal research questions from customers by telephone.

[8] Categories have been reordered to place similar categories along the same horizontal line.

Table II. Database Selection Query Categories—Domain Expert Responses

Domain Expert #1	Domain Expert #2	Domain Expert #3
⋆ Specific publications (journals, texts, magazines, etc.)	⋆ Sources or publications	⋆ Magazine or newspaper title ⋆ Reporter (bound case law docs) ⋆ Court/agency opinions (e.g., fed/state, int'l) ⋆ Statutes or codes (e.g., fed/state, foreign) ⋆ Federal congressional materials ⋆ Secondary materials (i.e., law reviews, etc.)
⋆ Documents or databases from a particular provider (e.g., American Bar Association)		⋆ Publisher
	⋆ Legal issue	⋆ Issue
⋆ Documents from a particular body (agency, court, commission, etc.)	⋆ Government entities (courts & agencies)	⋆ Specific court/agency
⋆ Databases relating to a particular topic (environmental, labor, securities, etc.)	⋆ Legal practice or research area	⋆ A [topical] "key" number (id)
⋆ Documents regarding a particular entity or person	⋆ Company name ⋆ Person name	⋆ Organization ⋆ Person ⋆ Group of people ⋆ Lawyer records
⋆ Databases relating to a geographic location	⋆ Geographic name	⋆ Place ⋆ State name ⋆ Foreign country
⋆ Citations to a particular document or set of documents (e.g., 1995 WL 303630, "Safe Water Drinking Act")	⋆ Document by citation ⋆ Document by title	⋆ Specific court/agency opinions ⋆ Specific statute sections
⋆ Specific database identifiers		⋆ A database name or id
	⋆ Definitions	⋆ A noun
	⋆ News & events	
	⋆ Financial queries	⋆ Statistics
⋆ Indeterminable	⋆ Category unclear	⋆ Unknown

remaining categories require some form of underlying metadata authority resource that can facilitate the mapping of user queries to their relevant sources of data (e.g., for courts/agencies and the aforementioned research/practice areas,[9] as well as geographic regions/locations.) Such metadata authority resources are

[9] There are roughly 50 major "practice" or "research" areas that are referred to in the legal domain (see Appendix B).

discussed in more detail in Sections 3.3 and 3.4. The remainder could benefit from existing collection selection techniques using searches run against, for example, a language model of terms and concepts present in a repository. This approach would be possible, for instance, for news or financial categories.

In order to be able to handle such a diverse set of queries, we investigate an interactive model that would invite users to participate in the selection of relevant collections or sets of collections. Once users assist with a basic categorization of their information need, the environment would provide access to desired data sets. The model would subsequently exploit characteristics of the incoming query, including language, granularity, domain, region, and other attributes that are typically ignored—or at the very least not explicitly exploited—by traditional information retrieval systems.

We have developed techniques motivated by actual user queries and their observed categories. These techniques permit users generating a spectrum of queries to simplify collection selection. By requesting category-type information along with queries, we have been able to implement this model in a large production environment without the need for massive metasearches and expensive metacollection builds and updates.

3.2 Addressing Query Categories

The baseline system in our experiment uses WIN's automatic selection and ranking of the top 20 collections and makes no differentiation between query types (Section 5). It runs all queries against a single database consisting of collection profiles, constructed by extracting top-level collection information from a database of collection content descriptions and other generic user subscription information. By contrast, the new system handles eight categories of queries:[10]

1. Sources & Publications;
2. Courts & Government Agencies;
3. Legal Practice & Research Areas;
4. Geographic Regions & Locations;
5. Legal Issues;
6. News;
7. Definitions; and
8. an "Other" category.

For our eight primary query categories, we use one of four distinct approaches (Figure 1).[11] Half of our methods rely on search; the other half rely on

[10] In this report, we do not treat query categories occurring significantly less than 2% of the time. In addition, our system has parallel and independent mechanisms for recognizing and handling queries with person and company names, and legal document citations. We do not address these in the remainder of the article, as they are separate query types.

[11] Figure 1: Regarding "Parallel Systems" on the bottom of the flow chart, when users have a bona fide document reference (e.g., *Roe v. Wade* or *142 Cal.App.2d 575*) or a person or company name, the system presents clearly visible options for directing the query to one of the parallel systems' entry points; L.M. in the box labeled (4) refers to a language modeling approach.

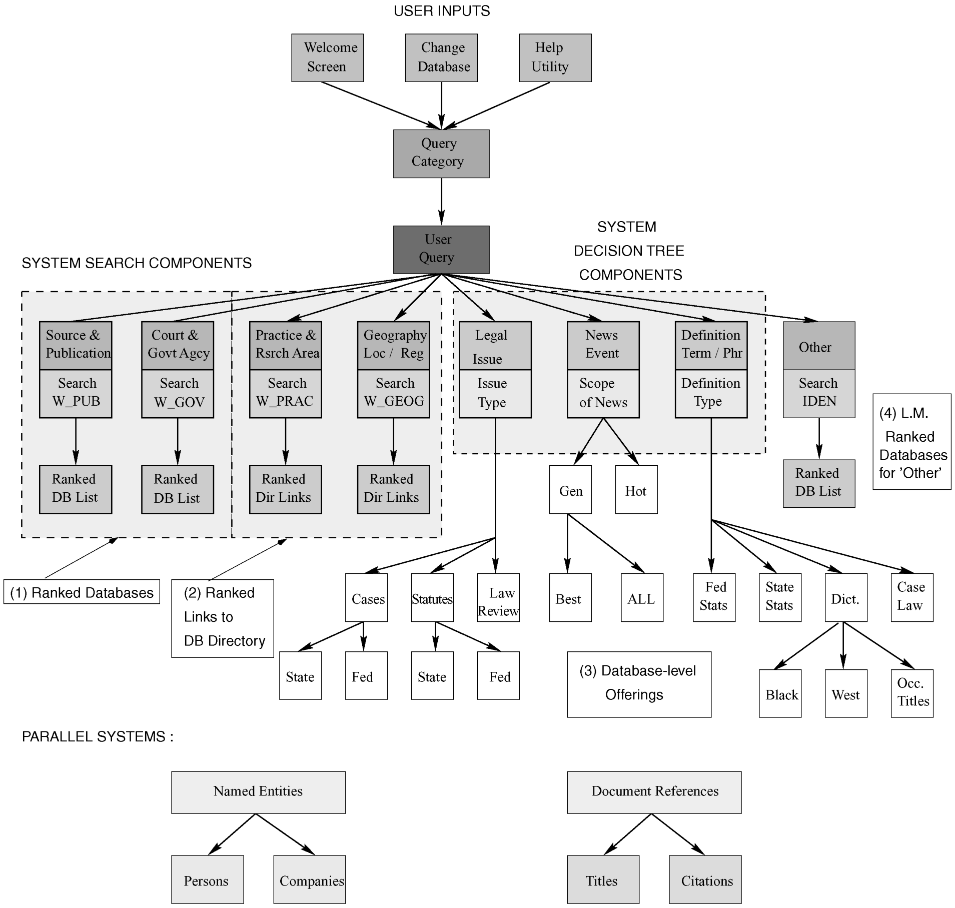


Fig. 1. Flowchart of preliminary operational system.

navigation, by using attenuated decision trees (e.g., Figure 2). The four methods invoking search run WIN against a category-specific metadata authority resource. Of these searches, the results are handled in two different ways, depending on query type: for the two largest authority resources, W_PUB (for publications) and W_GOV (for government agencies and courts), (with the finest granularity document profiles) actual *collection ids* are returned. The user can select and subsequently search one or more of these id-specified collections. For the other two authority resources, W_PRAC (legal practice areas) and W_GEOG (geographical locations), *directory links* to a topical or regional *subdirectory* are returned, to avoid presenting the user with flat lists of results (collections) consisting of several hundred individual collection descriptions. In these instances, the user can select the link and enter into a hierarchically organized directory in which to browse and find relevant collections. In the instances where search is not performed, the user is able to dig down into a simplified decision tree to find the most relevant set of collections within two or three levels [i.e., with respect to issues, definitions, news (e.g., Figure 2)]. In the decision tree mode,

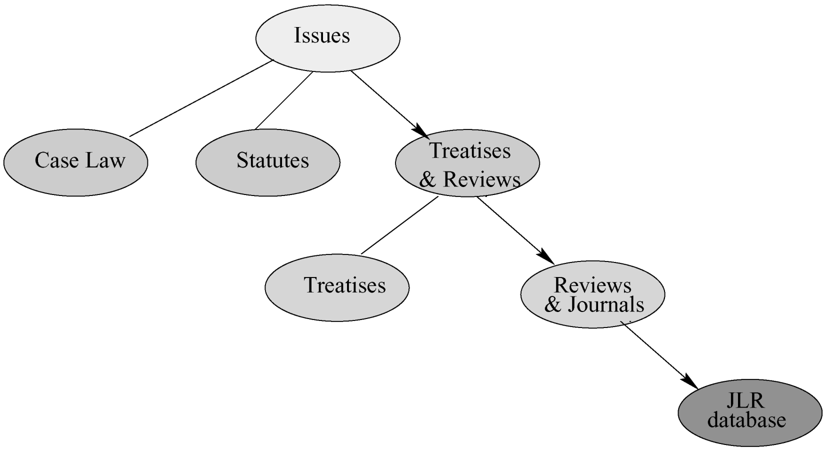


Fig. 2. Sample traversal for the issues category.

each path terminates with a large collection into which an assortment of important databases are bundled and where virtually all relevant related materials are found (case law, statutes, dictionaries, composite news, et al.). The "Other" category uses an approach analogous to language modeling of profiles for all of our collections.

The motivation behind using a specific approach for a given query category is based on the specificity of results a system could deliver for a given query category, where specificity is directly proportional to the granularity of the category profiles. For sources and publications (15,042 profiles) and courts and government agencies (3287 profiles), lists of top-ranked collections would permit a user to directly submit queries to one or more relevant databases. For legal practice and research areas (1352 profiles) and geographic locations and regions (300 profiles), knowing the desired practice area or region is still insufficient to know what document types a user is seeking (e.g., whether a user is looking for judicial opinions, statutes, or law reviews). Consequently, the most logical approach is not to deliver documents to these users but to deliver the user to the documents, that is, to provide them with links to the relevant portions of the Westlaw database directory, either for practice areas or geographic locations, depending on the query type. Lastly, for legal issues, news and current events, and term and phrase definitions, the deliverable options are simplified, thus permitting the user to navigate to the most relevant data source through a reasonably sized attenuated decision tree.

In the standard collection selection model, it is assumed that one does not have the resources to search each complete collection. Instead, one searches an index of collections, whether histogram-based or based on other metadata, obtains a ranking, and then searches the top-ranked collections for the most relevant documents. This would occur at the potentially serious expense of recall. In contrast, we propose interacting with the user earlier in the retrieval process in order to obtain greater confidence about the collections that merit further inspection.

3.3 Data

The metadata authority resources that support our searchable categories refer to specialized sets of database profiles. Each corresponds to one of the metalevel categories discussed above. Whereas Buckland et al. [1999] developed an Entry Vocabulary Technology to assist users in mapping their query vocabulary to that of potentially unfamiliar metadata vocabularies, we have developed "authority resources" around specific categories that professional users in the legal domain conventionally reference. They are designed to provide useful and effective matches with incoming user queries by focusing on specific taxonomies (e.g., legal practice areas). Rather than have one metadata repository containing database profiles for virtually every incoming query type, we have designed four indexable and searchable authority resources, each one focusing on a separate and distinct metalevel category that supports users' information needs. These include the following.

1. **W_PUB** (48.2%)[12]—maps user source/publication query to source/publication-related databases.
Profiles contain title of source or publication; alternative titles, alternative descriptions, related acronyms and abbreviations, and other domain-related descriptors.

2. **W_GOV** (7.5%)—maps user court/government agency query to appropriate district, state, or federal databases.
Profiles contain complete listings of US courts and government agencies and the database(s) where this court/agency material can be found.

3. **W_PRAC** (7.2%)—maps user legal practice/research area query to a database *directory* where related materials can be found.
Profiles contain listings of approximately 50 legal practice/research areas and links to their location(s) in a master (Westlaw) directory hierarchy.

4. **W_GEOG** (2.8%)—maps user location/region query to a database *directory* where related geographically related materials can be found.
Profiles contain listings of geographical locations/regions and links to the location of their associated materials in the master (Westlaw) directory hierarchy.

—**TOTAL** (65.7%)—Cumulatively, these authority resources treat two-thirds of the query types entering the DBS environment. Remaining query types are treated by simplified decision trees where, based on the type of legal issue, or definition, or news-related story, the user can navigate down a path to narrow the scope of the search to the relevant query-satisfying database (e.g., Figure 2).

3.4 Authority Resource Construction

In this research, we produce the four authority resource data sets described above and one general (baseline) data set of collection profiles, known as IDEN. Characteristics of the first four are described further below and in Table III. IDEN, by contrast, is a general source identification data set and is comparable

[12] Figures in parentheses refer to percentage of overall DBS queries (from Table I).

Table III. Collection Statistics for Metadata Authority Resources

Category	Data Set	Profiles	Size	Indexed Terms	Min./Max. Profile Length	Mean Length (Std. Dev.)
Source/ Publications	W_PUB	15042	2.34 MB	219817	11–6685	15 (94.7)
Court/ Gov't Agencies	W_GOV	3287	733 KB	91532	11–4747	28 (115.1)
Research/ Practice Areas	W_PRAC	1352	139 KB	11725	7–10	9 (2.6)
Geog. Regions/ Locations	W_GEOG	292	60 KB	4825	4–375	17 (27.8)
Legal Issues	Decision Tree—Primary and Secondary Legal Databases					
Definitions	Decision Tree—Definitional Sources					
News/Events	Decision Tree—Composite News Databases					
Other	Term-Based Collection Selection					

Table IV. Metadata Repository—Fielded (Partial Listing)

No.	Metadata Field	Data Type	Description, Examples
1.	Database Title	Text (short)	Complete Title of Database
2.	Database Identifier	Token (12 char)	Database ID or "Sign on"
3.	Coverage Dates	Date Range	Starting (Ending) Date of Publication
4.	Data Summary	Text (keywords)	Description of Docs Found in Database
5.	Subscription Options	Plan Type	Flat Plan, Hourly, Transactional
6.	Document Types	Class (of 50)	Category of Document
7.	Publication Types	Pub. Type	Electronic vs. Print
8.	State, Territory, Dist.	Region Type	States, Territories, or Wash., D.C.
9.	Jurisdiction or Nation	Juris. Type	Controlling Court or Applicable Country
10.	Nonlegal News Info.	News Type	Information Type
11.	Legal Practice Area	Prac. Type	E.g., Antitrust, Civil Rights, Taxation
12.	Multibases	DB ID List	Larger Databases Given DB is Contained in
13.	Information Provider	Publisher	Source of Data (e.g., Dialog, Dow Jones, West Group, etc.)
14.	Secondary Provider	Publisher	Original Source of Data
15.	Agency Name	Agency Id	Name of Government Agency or Similar
16.	Institution	Publisher	Law School, Pub. of Law Review, Journal,
...	...	...	...

to a verbose version of W_PUB, one that includes additional somewhat esoteric information about data provider and available subscription packages. The fields contained in the authority resources (database profiles) are automatically mapped from database records used for internal data management and maintained in a large relational metadata repository (Table IV). Related key concepts may be added to these fields, when judged useful, following human (paralegal) inspection. A majority of this fielded information is also contained in IDEN, but in a free-text format. This internal repository is not available for end-user searching.

Over 15,000 databases were used in these experiments, although not all were represented in each authority file due to the coverage of their associated categories.

```
<doc_no>189</doc_no>
<doc_id>AFRICANEWS</doc_id>
<title>Africa News</title>
<source>Dialog</source>
<lang>English </lang>
<descript>African News Services</descript>
<pubs>AllAfrica Press Service [All Africa]</pubs>
<pubs>Panafrican News Agency [Pan African]</pubs>
<pubs>Addis Tribune</pubs>
<pubs>Mail and Guardian of Johannesburg</pubs>
<pubs>Nigeria's Newswatch [News Watch]</pubs>
<pubs>Sudan Democratic Gazette</pubs>
<pubs>The Post of Zambia</pubs>
... ... ...
<loc>INT, AFR</loc>
<end_ref> ... </end_ref>
```

Fig. 3. Facsimile news collection profile (for publications).

```
<doc_no>5593</doc_no>
<doc_id>GLBLGOVERN</doc_id>
<title>Global Governance</title>
<title_exp>A Review of Multilateralism and International Organizations</title_exp>
<source>Dow Jones Interactive</source>
<lang>English </lang>
<multibases>ALLNEWS, MAGSPLUS, ENVNEWS, INTNEWS</multibases>
<descript>Economic Development</descript>
<descript>Human Rights</descript>
<descript>Environmental Preservation</descript>
... ... ...
<title_src>Economic Development</title_src>
<loc>INT, ASA, AUS, CAN, EUR, NZ, US, UK</loc>
<end_ref> ... </end_ref>
```

Fig. 4. Facsimile international review collection profile (for publications).

The four primary authority resources concentrate on publication and government (collection-based), topical and regional (link-based) paths to data, to access available collections. Simplified facsimile samples of W_PUB publication profile "documents" are shown in Figures 3 and 4 and a W_GOV profile document in Figure 5. W_PUB contains one document profile for each searchable collection in the system. Its construction was thus the most straightforward of the four. W_GOV contains one document profile for each court/agency or set of courts represented in collections in the system, and is thus slightly less granular. Its construction required additional filtering and merging of court-related information stored in the master metadata repository. It took a paralegal approximately two weeks to complete. W_PRAC and W_GEOG are less granular still and represent links to *sets* of collections organized by topic and region, respectively, in the Westlaw database directory. Because there are roughly 50 legal practice areas and these are also recorded for each applicable database in the master repository, the construction of W_PRAC took about one week of paralegal time. Lastly, W_GEOG contains only several hundred entries, each

```
<doc_no>7239</doc_no>
<doc_id>ENFLEX-LA</doc_id>
<title>Louisiana Environmental, Health and Safety Regulations</title>
<source>IHS Environmental</source>
<lang>English</lang>
<multibases>ENFLEX-STATE</multibases>
<agency>Louisiana Department of Environmental Quality</agency>
<alternate>Louisiana Department of Natural Resources</alternate>
<alternate>Louisiana Department of Public Safety</alternate>
<alternate>Louisiana Environmental Control Commission</alternate>
... ... ...
<loc>US</loc>
<end_ref> ... </end_ref>
```

Fig. 5. Facsimile environmental, health, and safety regulations profile (for courts and government agencies).

one geographic in nature and containing a link to the various regions' materials in the Westlaw database directory. Its construction took a paralegal less than one week to complete. The scope of each of these data sets explains the appreciable difference in size between the *collection-based* authority resources (W_PUB and W_GOV) and the *link-based* authority resources (W_PRAC and W_GEOG), and the inverse relationship between authority resource size and associated granularity. As an illustration, W_PUB is clearly the largest authority resource, yet possesses the smallest granularity. By contrast, W_GEOG is the smallest authority resource, but it has the largest granularity.

Updates to the source and publication authority resource are performed automatically. When a database is added to the Westlaw system, new profiles are generated from the master repository. These profiles are reviewed for completeness, however, by a domain expert. Authority resources developed around government institutions (courts and agencies), legal topics (practice and research areas), and geographic topics (locations and regions) are generally more stable; thus updates to these resources are minimal. For instance, when a practice area in W_PRAC becomes outdated (e.g., Y2K) or a new practice area appears (e.g., digital copyright), associated profiles are removed or created in a semiautomated manner, under minimal paralegal supervision. The same would apply to government agencies in W_GOV or nation-states in W_GEOG. In the case where similar authority resources would be developed for another national jurisdiction, for instance, for Canada or Australia, the processes would be the same, although the work effort would be reduced since the scope and corpus of documents would not be of the same magnitude in our system as that for the US.

It is worth pointing out that the total number of collections represented in this research does not correspond to the sum of the collections profiled in the four authority resources. In reality, the four authority resources provide alternative *views* of the collections, each using a different category-specific perspective. Since there is a one-to-one correspondence between source/publications and databases, W_PUB represents the cardinal number of collections represented: 15,042. By contrast, W_GOV, W_PRAC, and W_GEOG provide alternative and less fine-grained characterizations of the same collections.

3.5 The Role of User Subscriptions

In very large proprietary data environments like West Group's and Dialog's, many thousands of online databases may be available to clients. It is thus common for clients with different sized enterprises and with different information needs to have different subscription arrangements to cover the cost of accessed information. Some subscribe on a transactional pay-as-you-go basis, some choose unlimited access to a small number of select, relevant databases (e.g., within their jurisdiction or practice area), and still others have basic coverage plans with the option to expand access on a per-need basis. Subscription arrangements and the variability of cost are one reason why database selection can be a two-step process for users unfamiliar with available resources. Accordingly, clients play an important role in deciding on the scope of their research, based on perceived relevance, value, and cost of accessible materials.

3.6 Indexing

For each of the authority resources, metalevel information is maintained in XML-like tag sets. To aid retrieval, the majority of these tagged elements are indexed although some are not. Fields not indexed might include those that contain concise text strictly for presentation purposes or information to facilitate internal organization and classification of collection profiles.

Examples of these metalevel profiles are shown in Figures 3 through 5. Fields in these profiles that are not indexed include associated title fields (used for presentation only) as well as multibase (CONTAINED-IN) fields. Other fields are indexed and virtually all indexable fields are first stemmed. Stemming is performed as it is not uncommon for users to enter variations of titles or descriptive terms for publication (e.g., *AIDS Therapy* instead of *AIDS' Therapies*), court (e.g., *State of New York Court of Claims* instead of *State of New York Courts of Claims*), or practice areas (e.g., *Commodity Regulators* instead of *Commodities Regulation*). We use the Porter stemmer with a stopword list of approximately 300 common terms.[13]

3.6.1 *Special Considerations for Legal Collections.* It is worth noting that a standard case law opinion, to take one example of a legal document, is typically two to five times the length of a Web document (30 to 50 KB vs. 10 to 15 KB) [Murray and Moore 2000; Heydon and Najork 1999; Conrad and Dabney 2001]. Although there are circumstances in which we can and do use language model type term distribution histograms to represent the vocabulary of a collection, when tens of thousands of collections are available, language models may not be the most effective approach to collection selection. That is, some collections possess language very similar to that of adjacent collections, whereas others are subsumed by larger "multibase" collections (e.g., *Minnesota Environmental Statutes* are contained in *All States Environmental Statutes*). This hierarchical relationship is another reason why it can be useful for clients to play a greater

[13] In our standard production environment, however, virtually all terms are indexed for purposes of specific title or contextual retrieval.

Sources/Publications:
(1) Occupational Safety & Health Review Comm. Decisions
(2) Venture Capital Journal
(3) Pennsylvania Insurance Department Records
(4) Computer World
Courts/Government Agencies:
(1) California Railroad Commission
(2) US District Court for the Southern District of NY
(3) Voluntary Labor Arbitration Tribunal
(4) Dept of Transportation Coast Guard Merchant Marine
Practice/Research Areas:
(1) Juvenile Justice, Child Welfare
(2) Professional Responsibility and Ethics
(3) Fair Employment Practices
(4) Patents and Trademarks
Geographic Locations/Regions:
(1) District of Columbia
(2) Sonoma County
(3) England or British
(4) West Indies

Fig. 6. Sample database selection queries by category.

role in the steps that narrow their candidate collections when database selection is performed (see Example, Appendix C).

3.7 Test Queries

For research and testing purposes, we use two sets of 450 actual user queries, further broken down by category. Each of the two sets of 450 queries originates from a different month's database selection query log. Each of these logs contains queries that are assigned metalevel query categories corresponding to those in Table I. They originate from a WIN-based database selection application (IDEN). The queries were randomly selected, with sufficient numbers chosen to comprise the categorized query sets of 50 or 75 that are reported. To improve the accuracy of our evaluation measures, we required each of our query sets to contain at least 50 real user queries [Buckley and Voorhees 2000], and for the purposes of comparing variance, to have at least two query sets for each of the categories we inspected. We call these paired sets A and B. Furthermore, our W_PUB query sets are 50% larger than those for the other categories since our analysis showed that nearly 50% of all database selection queries appeared under this classification. The categories we use in these experiments include (1) sources/publications, (2) courts/government agencies, (3) practice/research areas, and (4) geographical regions/ locations. In all, we have 200 queries per category (300 in the case of sources/publications) or 900 total. The four categories are further divided into subsets of 50 queries each (75 each in the case of sources/publications; see Table IX). We run each of these sets against authority resources indexed for WIN-based retrieval. Samples of these queries can be seen in Figure 6. Queries not falling into our most frequently occurring categories, that is, falling under the "Other" category, are used in a standard collection selection test run which is beyond the scope of this article. Average

Table V. Query Categorization: Expertise Among Legal Practitioner—Participants[a]

Expertise Matrix	L.S. Experience	L.S. Experience
D.B.S. Experience	*Attorney*	*Attorney*
D.B.S. Experience	*Paralegal*	*Attorney*

[a] D.B.S. = Database Selection; L.S. = Library Science.

Table VI. Kappa Statistics for Categorization Performed by Four Assessors[a]

Tokens per Query	No. Queries	Kappa Statistic	Associated z	Significant
1	40	$\kappa = 0.7697$	20.15	Y
2	40	$\kappa = 0.7220$	16.50	Y
3	40	$\kappa = 0.8022$	7.70	Y
4	40	$\kappa = 0.9106$	17.66	Y
5 or more	40	$\kappa = 0.9117$	16.06	Y
Combined	200	$\kappa = 0.8232$	38.98	Y

[a] [$z = 2.32$ for $\alpha = 0.01$]; $\kappa = 1$ for complete agreement among assessors; $\kappa = 0$ for no agreement among assessors.

query lengths vary from four to seven terms for W_PUB and W_GOV to one to three terms for W_PRAC and W_GEOG.

3.7.1 *Validation of Query Categorization.* Because these queries were initially categorized by an attorney who is familiar with our metalevel query categories, our results should be presented as upper bounds on expected performance. Nonetheless, to investigate how reasonable it is to expect users to categorize queries reliably, we performed a validation experiment that involved four individuals with four different levels of legal training. They included one paralegal and three attorneys. Of the attorneys, two had no familiarity with database selection queries; one did. In addition, two of the attorneys had training in library science. In short, these subjects represent fairly well the spectrum of legal practitioners who use a system like Westlaw (Table V). In this experiment, each of the participants was given a set of 200 real user queries and asked to categorize them using the first 12 categories shown in Table I. The queries were randomly selected from a single week's query log. In order to avoid any particular category and its inherent length from dominating the results, enough queries were selected so as to permit the generation of five subsets of queries, each set based on a different query length (e.g., with number of tokens = 1, 2, 3, 4, and 5 or more tokens).

The only pretrial inspection that was made was to determine whether the query would be reasonably interpretable by someone with some degree of legal training. Hence a query such as "alsdkjf" would be discarded. We used the Kappa statistic for nominally scaled data to compare our interassessor concordances for the 200 queries and the 12 categories [Siegel and N. John Castellan 1988]. We explored interassessor agreement relative to query length since it has been shown that longer query statements reduce the ambiguity associated with very short queries [Sanderson 1994]. We wanted to determine if this same relationship would hold true for this task as well. The results of our comparisons are presented in Table VI.

Computational linguists have taken $\kappa = 0.8$ as the norm for significantly good agreement, although some argue that there is insufficient evidence to choose 0.8 over, for instance, other values between 0.6 and 0.9 [Marcu 2002]. To underscore the significance of the above values, it may be useful to mention that of the 200 queries, the four assessors were in complete agreement on 158 of them and three of the assessors agreed on an assignment for 28 others.

The concordance between the original domain expert's classifications and those of the four assessors above was also determined for the 200 queries (to represent the assessors' category for a given query, we used the category upon which a majority of them agreed).[14] This categorization comparison resulted in a kappa statistic of $\kappa = 0.9196$. The four assessors' majority category agreed with the original domain expert in 185 out of the 200 queries.

These results, together with the fact that the kappa scores tend to monotonically increase with token length (one-token queries excepted), illustrate that the longer the query, the more concordances one can expect among different "users." So in this application as well, there appears to be a relationship between query length and query clarity or disambiguation [Sanderson 1994].

3.7.2 *Testing the Significance of the Kappa Statistic κ.*

After determining the value of the kappa statistic κ, it is customary to determine whether the observed value is greater than the value that would be expected by chance. This can be done by calculating the value of the statistic z, where,

$$z = \frac{\kappa}{\sqrt{var(\kappa)}}$$

in order to test the hypothesis $H_o : \kappa = 0$ against the hypothesis $H_1 : \kappa > 0$ [Carletta 1996; Siegel and N. John Castellan 1988].

The above value of κ for the combined query set yields $z = 38.98$. In addition, the value of κ found when comparing the original domain expert to the four assessors yields $z = 95.58$. These values exceed the $\alpha = 0.01$ significance level (where $z = 2.32$). Therefore, we may conclude that the assessors exhibit significant agreement on this categorization task. (See Table VI above for the corresponding z for each query length subset.) These results suggest that a group of legal practitioners with diverse expertise are capable of categorizing their information needs with a considerable degree of similarity. These results in turn mean that it is reasonable to expect that most users of this feature would be able to choose the "correct" category in which to continue their database selection search.

3.8 Relevance Judgments

The relevance judgments used in these experiments are made in response to test runs on the four searchable authority resources, each associated with a different category of query (as reported in Sections 3.3 and 3.4). The judgments are binary in nature (i.e., relevant/not relevant) and are performed by one attorney with a graduate degree in library science.

[14] The number of ties among the assessors was negligible. In these cases we assigned the assessors' majority category the one that disagreed with the domain expert's categorization.

Because our users place a premium on precision at top ranks, we focus special attention on the top ranks in which relevant collections appear. We report whether relevant results are in the top 5 as well as the top 20 ranks and if those results include one, some, all, or none of the relevant collections available. We place this emphasis on the top 5 collections because users looking for relevant databases with which to begin their research have little patience to examine 19 of 20 candidates before encountering the first database containing relevant documents. This top 20 analysis and evaluation is performed on a total of 900 queries (i.e., two groups of 450 queries, Section 5.1). In the vast majority of cases, the query types that invoke search can achieve very high collection recall in the top 20 results (with values surpassing 90%, due to the specificity of many queries). This evaluation thus permits us to address both precision and, implicitly, recall for our user query sets.

Our recall estimates are based on information supplied by domain experts who are intimately familiar with the collections. Unlike the pool of judgments in the TREC environment, we did not have a large, existing set of relevance judgments available at the start of these trials [Voorhees and Harman 2000]. Because of the nature of our relationship with the sponsoring department, we could not ask for unlimited judgments for all 15,000 databases for each query. We were thus required to construct our own sets of judgments relying on the contributions of legal database domain expertise. This approach was viewed as reasonable from both a practical and evaluative standpoint. Based on their knowledge of controlling jurisdictions, appropriate legal practice areas and applicable document types (e.g., judicial opinions, statutes, law reviews, etc.), these experts are skilled in reducing the set of potentially relevant material to a relatively small percentage of the overall number of collections available. So the expert's judgment is believed to be fairly reliable. We acknowledge that our pool of positive relevance judgments is almost certainly a subset of the complete set of positive relevance judgments, but it is likely a very large subset. Given our domain expert's years of experience with law and relevance assessments, we believe that this background at least partially mitigates concerns over the degree of bias present in the judgments supporting our recall evaluation. We thus contend, as does TREC [Voorhees 2002], that our recall estimates are close enough approximations to be useful when comparing systems and, in our case, to permit the identification of significant omissions indicative of more serious problems with the search strategy.

In addition, our results are compared with an existing system (IDEN) that processes all queries in the same manner by using one database of collection profiles containing titles and a variable-length, free-text description of contents (see Section 3.4). These were judged for relevance in the same manner. This latter comparison was performed using the first set of 450 queries (Section 5.2).[15]

[15] The second set of 450 queries was not evaluated for comparative precision performance because the domain expert providing judgments for our results was reassigned to work on another project near the end of our evaluation process.

Table VII. Relevance Judgment Concordances

Query Category	No. of Queries with Complete Agreement	No. of Judgments in Agreement
Source/Publication	10/20 (50%)	304/351 (86.6%)
Court/Gov't Agency	6/10 (60%)	160/167 (95.8%)
Legal Practice Area	8/10 (80%)	133/135 (98.5%)
Geographical Location	10/10 (100%)	10/10 (100.0%)
Combined	34/50 (68%)	607/663 (91.6%)

Table VIII. Significance Tests for Relevance Judgment Concordances

Query Categories	No. Queries	No. Judgments	Sign Test ($N = 16$)	Wilcoxon Signed Ranks Test ($N = 16$; $\sum_{q=1}^{n} \lvert \mathit{Diff}_w \rvert = 56$)
Combined	50	663	($N^+ = 8$; $N^- = 8$)$\Rightarrow$ $\mathbf{H_o}$	($\mathit{Diff}_w^+ = 17; \mathit{Diff}_w^- = 39$)$\Rightarrow$ $\mathbf{H_1}$

3.8.1 *Interassessor Validation.* In order to perform a preliminary investigation into the dependability of these judgments, a second attorney with no background in library science participated in an interassessor study. The second attorney was intended to represent a conventional legal practitioner—user. The study was conducted by means of an experiment in which the pair of attorneys provided relevance judgments for databases returned in response to 50 actual user queries. The queries included four sets of 10 queries, one set for each authority resource, plus an additional 10 from the most dominant category, sources and publications (W_PUB). All were randomly selected from their query category. Judgments were made on the top 20 databases returned. For some queries, less than 20 databases were returned, especially in the case of geographical locations where fewer possible matches can occur. In this interassessor correlation study, the two attorneys assessed the results and were in agreement for 92% of the queries (Table VII).

The second column in the table represents the percentage of queries for which the assessors were in complete agreement.

3.8.2 *Testing the Significance of the Interassessor Concordances.* The null hypothesis for these interassessor consistency tests is that there exist no significant differences between the relevance judgments made by the two judges. The sign test provides little evidence against the null hypothesis insofar as there were 8 out of the 50 queries for which the first assessor produced one or more positive relevance judgments than the second assessor and there were 8 queries for which the second assessor produced one or more positive relevance judgments than the first (Table VIII). By contrast, the Wilcoxon test favors the alternative hypothesis (one of the assessors will produce more positive relevance judgments than the other) inasmuch as the magnitude of the differences is significant. This is attributable to two queries for which the difference in positive judgments between the two judges is greater than 3 (out of a total of 20). In general, the assessors' judgments matched for 92% of the collections they judged (Table VII). It was the second assessor, without the library science background, who tended to cast the net more broadly, and who thus favored recall. By contrast, the first assessor, with the library science background, appeared to exercise a finer definition of relevance, thus emphasizing "on point" collections. The average additional positive relevance judgments for the two queries in question was 13. The two queries were "Combined Federal and State Cases" (which has numerous candidates to choose from, both complete and partial) and "Dun & Bradstreet" (which also has numerous choices, both US-based, and non-US-based). The second assessor gave more positive relevance judgments to results from these two queries because of the larger amount of tangentially relevant material.

It is important to underscore that result sets produced by the two assessors were never involved in intersystem comparison. Rather, the first assessor's judgments were exclusively used to evaluate queries from both months as well as queries run against the baseline system [Sections 5.1 (Table IX) and 5.2 (Table X)]. Moreover, the localized differences in judgments between the two assessors should not be viewed as significant from a system evaluation point of view because there is evidence that "*comparative* evaluation of retrieval performance is stable despite substantial differences in relevance judgments" [Voorhees 1998]. Were their concordances not in the 90% range, we may have opted to invest more human resources in query examination. Yet such a reallocation in resources may have resulted in a reduction in the size of the query sets due to the cost of domain expert participation. Ultimately, it may have meant compensating for one potential deficiency by introducing another.

4. COLLECTION RANKING

Much collection selection research is based on applications of IR techniques to the distributions of terms and phrases that comprise collections. It is assumed that the statistics that characterize collections are readily available or can be approximated through the iterative use of probing queries [Callan and Connell 2001]. It is also assumed to be too costly to query all the available collections, so a restricted number are selected based on some fixed threshold derived from a score or a fixed number of collections.

INQUERY's and WIN's algorithms for ranking documents have been previously reported [Turtle 1991; Broglio et al. 1993; Allan et al. 1997]. In this database selection application, the document retrieval model is used since we are working with condensed representations of the collections (i.e., collection profiles similar to those in Figures 3 through 5). $tf \cdot idf$ scoring is applied to calculate the probability of relevance or belief score for a given collection profile p_{bel}.[16]

$$p_{bel}(w_i|c_j) = d_b + (1 - d_b) \cdot tf_b \cdot idf_b,$$

[16] This version of $tf \cdot idf$ scoring has been implemented in the Westlaw production environment. As document length has not normally been stored with a document's metadata, scoring formulae such as BM25 have thus far not been implemented [Robertson et al. 1995].

where

$$tf_b = d_t + (1 - d_t) \cdot \frac{\log(tf_i + 0.5)}{\log(tf_{\max} + 1.0)}$$

$$idf_b = \frac{\log\left(\dfrac{N + 0.5}{n}\right)}{\log(N + 1.0)};$$

n represents the number of collection profile documents in which the query term w_i appears and N is the total number of collection profile documents. d_b is the minimum belief component and d_t is the minimum term frequency component when the term t_i is present in a collection representation c_j.

We use a reduced stop word list because of the role certain common words can play in titles and title descriptions. We also use a standard Porter stemmer [Porter 1980]. In addition, we rely on distilled consonant representations of terms present in the authority resources. The latter help determine matches when users make common spelling errors or invoke nonstandard abbreviations. We further apply query expansion techniques using a domain-specific thesaurus and acronym expansion, as well as other word forms when individuals use numerals or other terms with common or reasonable synonyms (e.g., *code* ⇔ *statutes*).

We have also found it necessary to modify the nature of WIN's combined document plus best passage scoring. Because of the nature of the collection representations, we use a modified document scoring formula. In a standard WIN search, the final score of a matched document would typically represent some weighted average of a whole document score and the best passage score. In our case, as evidenced by the sample collection profile shown in Figure 3, we equate the score for a given document with the score of the top-ranked passage, or in this case, top-ranked field (e.g., <pubs>). In this manner, if a publication or court query focused on a specific title or court, and the match occurred exclusively within a certain field in a document, that field's score alone would be promoted to represent the collection profile's score. This approach avoids incidental double term weightings because the document is subdivided into independent matchable entities. Thus a query focusing on "news on the postal system and mail delivery in Nigeria," would avoid incorrectly promoting the sample document in Figure 3 because of its many hits. Rather, only the highest scoring <pubs> field would be promoted in assigning the document's final score. Not all fields are segmented in this way, but certain key fields containing publication, court, or other similar listings do exhibit such divisions.

5. RESULTS

5.1 Individual Performance Evaluation

In the experiments described above, two sets of 450 category-specific user queries were run against the corresponding authority resources. Real user queries, originating from two separate months, were used for the task. The results for the query sets are shown in Table IX. Queries were obtained from

Table IX. Performance Evaluation: Precision—A. Month I (top) and B. Month II (bottom)

Category/	Class							Total
Test Set	1	2	3	4	4(a)	4(b)	4(c)	Queries
Publication A	42	11	6	16	8	1	7 (9%)	75
Publication B	43	8	5	19	12	0	7 (9%)	75
Court/Agency A	37	3	2	8	2	2	4 (8%)	50
Court/Agency B	26	15	1	8	1	2	5 (10%)	50
Practice Area A	40	5	1	4	0	0	3 (6%)	50
Practice Area B	41	7	1	1	0	0	1 (2%)	50
Geographic A	42	2	0	6	2	1	3 (6%)	50
Geographic B	39	0	0	11	2	1	8 (16%)	50
Total	310	51	16	73	27	8	38	450
Percent	68.8%	11.3%	3.6%	16.7%	6%	2%	8%	100%
Percent, excl 4(a)	73.3%	12.1%	3.8%	11.3%	—	2%	9%	423

Class 1. Single or all relevant source(s) in ranks 1 to 5.
Class 2. Most of relevant sources in ranks 1 to 5,
or single relevant source in ranks 6 to 20.
Class 3. Some of relevant sources in the top 20 ranks presented.
Class 4. No relevant sources in ranks presented
4(a) Not available in system
4(b) Difficult to match without further information
4(c) Source in system but not presented (missed).

Category/	Class							Total
Test Set	1	2	3	4	4(a)	4(b)	4(c)	Queries
Publication A	49	5	2	19	10	1	8 (10%)	75
Publication B	50	3	1	21	12	1	8 (10%)	75
Court/Agency A	36	1	2	11	8	0	3 (6%)	50
Court/Agency B	37	6	0	7	1	3	3 (6%)	50
Practice Area A	41	3	1	5	2	2	1 (2%)	50
Practice Area B	43	4	2	1	0	1	0 (0%)	50
Geographic A	49	0	0	1	1	0	0 (0%)	50
Geographic B	47	2	1	0	0	0	0 (0%)	50
Total	352	24	9	65	34	8	23	450
Percent	78.2%	5.3%	2.2%	14.4%	8%	2%	5%	100%
Percent, excl 4(a)	84.6%	5.8%	2.4%	7.5%	—	2%	6%	416

two distinct months to determine whether there exist significant variances over time. The results presented are as much qualitative as they are quantitative. We have determined that for systems that place a premium on precision at top ranks, broadly defined relevance classes may more clearly indicate performance differences between query categories (definitions for these relevance classes are located in Table IX). In addition, the domain expert involved in evaluating these results had extensive familiarity with our data and knowledge of which results would be considered acceptable to representative users, given sufficiently broad evaluation metrics.

The motivation for this nonstandard approach to evaluation was fourfold. First, we did not have a fixed set of queries with mature, preexisting relevance judgments, as in the case of some of the tracks at TREC conferences [Voorhees and Harman 2000]. Second, we also had at our disposal legal and library science domain expertise which contributed a solid grasp of the collection-level sources available, at least in response to the four types of queries corresponding to our searchable categories. Third, in the absence of a TREC evaluation environment and the more exhaustive resources it might take to produce one, we wanted to develop a set of qualitative relevance classes that would explicitly produce collection-level precision values, but also implicitly yield collection-level recall values. Finally, the domain expert had the latitude to perform online research, when helpful, to inspect more carefully the quality of a given result set, relative to a query, before making an assessment. This latitude gave the domain expert the opportunity to "get inside the head of the user" when assessing results. The classes used represent a hybrid of relevance and rank information in order to embody user-centered relevance indicators. The operational definitions that resulted can be found below. In each case the top 20 candidate collections are examined.

Classes:

1. *single or all* relevant sources are present in ranks 1 to 5 (highest-level precision); translates into 100% recall;
2. *most* relevant sources in ranks 1 to 5, or single or all relevant sources in ranks 6 to 20 (moderately high precision); translates into over 50% recall;
3. *some* relevant sources in the top 20 ranks (medium precision); translates into 50% or less recall;
4. *no* relevant sources in the top 20 ranks (lowest precision); translates into 0% recall.

The rationale for the Class 2 definition is as follows. Since Class 1 handles the case where all relevant sources appear in ranks 1 to 5, Class 2 handles the case where these sources are not in the top 5 ranks, but are nonetheless still in the top 20 ranks. In addition it includes the case where most (but not all) relevant sources are in ranks 1 to 5, thus preventing inclusion in Class 1, but still warranting inclusion in a moderately high precision class. In brief, Classes 1 and 2 make distinctions between two types of results sets. They distinguish between results with the single relevant collection in the top 5 from results with the single relevant collection not in the top 5. They also distinguish between results with several relevant collections in the top 5 from results with several relevant collections not in the top 5. This was done in the interest of preventing too much granularity from weakening any potentially meaningful conclusions. It is important to note, however, that these specific searchable categories tend to service queries that have a single on-point database or "answer," which is why a premium is placed on Class 1 results.

Our initial interest was in discriminating higher precision results (Classes 1 and 2) from lower precision results (Classes 3 and 4). For each of our query sets, between 85 and 90% of our queries produced results in Classes 1 and 2 (discounting queries for which no relevant sources were available). Although we were not able to investigate users' perceptions of the *quality* of their results, our domain expert was able to determine which collections possessed the highest probability of relevance for users' information needs.

Table X. Performance Comparison: Category-Based System Versus IDEN

Category/	Class (%)							Total	Total
Test Set	1	2	3	4	4(a)	4(b)	4(c)	(%)	Qrys
System	68.8	11.3	3.6	16.7	6.0	2.0	8.0	100	450
IDEN	26.3	14.3	40.3	19.0	6.3	2.7	10.0	100	450

One might have expected an appreciable degree of performance variation across query categories, yet the levels of precision examined across the four searchable categories do not reveal significant variance. With the exception of month I's Court and Agency sets A and B, there do not appear to be any appreciable differences between each of the query categories' sets A and B. It is also worth noting that no changes occur in the relative positions of the four classes over time when one compares the sizes of the result sets between the two months (for Classes 1 through 4). We have observed that these four categories tend to represent concrete rather than abstract concepts (e.g., publication *titles*, court *names*, specific practice *topics*, geographic *locations*). This is probably one reason why the system is usually able to capture the most salient database profiles in the top five ranks.

5.2 Baseline Comparison

In addition to category-specific precision figures, we compare results from the first month's set of 450 queries with results obtained from the baseline IDEN system. The same WIN retrieval method is used for both the IDEN system and the category-based system.[17] As IDEN references a more verbose version of W_PUB, its results are compared to results from queries run against W_PUB. These results, shown in Table X, suggest that category-specific precision for Class 1 (i.e., the top result class, single relevant source in top 5 ranks) is increased by nearly a factor of 2.5 (averaged over 450 queries). Most of the collections not correctly identified or promoted by IDEN to Class 1 show up instead in Class 3 (i.e., relevant source(s) farther back in the ranks). One simple explanation for the significant improvement in results for the query category approach is that a certain amount of collection filtering has taken place in the construction of these authority resources, as evidenced by the number of collection profiles present in these authority resources (Table III). Hence fewer nonrelevant collections are present that could dilute the performance of the candidates the system delivers (i.e., less than IDEN). The lesson we have learned is that by including the user earlier in the decision loop, it may be possible to eliminate subsequent iterations of user—system interaction.

Category-fielded query submission has been available in our production environment for over a year. If client usage is an indication of improved user access and system performance, the category approach greatly surpasses that of its predecessor. Usage of the new utility has increased to nearly 5000 queries per day. By contrast, the previous system, IDEN, averaged roughly 800 queries per day. More recently, usage of the primary authority resources, W_PUB and

[17] The only difference that exists between the two systems is that the new system marshals an expanded acronym expansion facility that was unavailable in the production IDEN system.

W_GOV, has placed them among the top 1% most used collections of the more than 15,000 available on Westlaw.

6. DISCUSSION

This early user interaction approach, leveraging categorized queries, is viewed as a complement to existing collection selection techniques. It harnesses certain special query types and has the ability to exploit them more effectively. The approach accomplishes this by offering a heightened role for users to assist the system in its selection. Such increased user participation serves as the backbone to a tool that fosters more user control in searches. In addition, a category-fielded system appears to accommodate concrete, detailed queries better than abstract or vaguely worded queries, at least with respect to those categories in which queries are run against dedicated authority resources. Our studies show that these specific queries represent two-thirds of all queries (Table I). Of the remaining third, queries of a more abstract nature would be directed to the remaining portion of the system, namely, to the "Issues" and "Other" paths. These involve an even greater degree of user participation and thus would introduce more subjectivity to any proper evaluations. We concluded that to be conducted in a valid and rigorous manner, such evaluations were beyond the scope of our resources. Thus the evaluations we performed focus on comparisons between real user queries run against a central baseline database of collection profiles and queries run against several category-based authority resources. Although the results are by no means definitive, and are intended to be viewed qualitatively as much as quantitatively, our results suggest that this approach can help eliminate the diluted precision that can often occur with conventional collection selection techniques. Furthermore, because of the specificity of many of the categorized queries, improved precision may be achieved without the expected compromise in recall, given that a complete result set is often found in the top 20 ranks.

This categorized query approach may be more appropriate for proprietary data environments like Westlaw or Dialog than for rapidly growing data environments like the Web. Such authority resources are clearly easier to construct for proprietary data environments where it is possible for human or automated resources to approach comprehensive knowledge of the scope and focus of most collections. Yet given that researchers have acknowledged the "black areas" of (unindexed) data that exist on the Web, it may be possible to focus on some of the most important resource areas of the Web as well—by leveraging this "authority resource" approach. At the very least, it would be possible to generate essential core terms or representations for document records in a Web environment from minimal metadata, even if one did not have complete access to an entire indexed database. This is where our approach has potential advantages over common database selection indexing methods.[18]

[18] Our approach also shares characteristics with the traditional library science approach in which researchers are directed to the appropriate type of resource (e.g., journal or dictionaries) based on the expression and analysis of their information need [Bopp and Smith 2000].

Another issue we have had to address involves the overlap of our query categories (Table I). A user-oriented system ideally needs to be sufficiently robust to deliver the same relevant collections to users regardless of the query category they select, as long as it is a reasonable selection. For this reason we have also tested our system using similar information needs entered through multiple category paths. For example, if someone enters "New Mexico Bankruptcy Procedures" via either the Source category (New Mexico) or the Practice Area category (Bankruptcy), they should expect to end up with similar paths to relevant collections; likewise, if someone enters "Canadian Environmental Law" through either the Practice Area category (Environmental Law) or Geographic category (Canada), they should encounter similar sets of relevant collections. We have found that the most relevant or "on point" collections do reliably appear in such scenarios. Differences do exist, however, with marginally relevant collections. Yet it is an open question what *utility* such marginal collections would actually contribute, given the high degree of scrutiny professional users exercise.

It would also be interesting to investigate user performance when providing only IDEN's collection ranking, in order to determine how useful its singular collection selection mechanism is, without the additional processing developed to support the categorized approach (described in Section 4). We are pursuing internal assistance to simulate such an evaluation.

7. CONCLUSIONS

We have shown that by permitting users to collaborate with an information finding system, users can achieve high-precision results effectively without the need for more computationally expensive mechanisms. Our approach is computationally inexpensive insofar as it relies on relatively modest authority resources that consist of databases containing on the order of tens of thousands of concise collection profiles. Such front-end handling can contribute significantly to the efficiency of large online systems with hundreds of thousands of users and tens of thousands of data sources.

The research presented in this article is novel in several respects.

—It is completely motivated by actual information needs expressed by users in the particular domain;

—In several instances, the assumptions made in the development of this query-category-based model have been validated through the direct involvement of domain experts, library scientists, and legal practitioners;

—This model attempts to bridge the divide that has long existed between computationally exhaustive systems deficient of any user—system interaction and information theories that stress the ongoing role of the user in search strategies.

Our results are not inconsistent with Park's [2000] findings: that more user input is important in large environments with distributed data, that distinct database characterization can assist users with choices for integration, and that certain users prefer control over their database selection processes. These findings also call into question the assumption that interaction with merged data is most effective. We view what have evolved into conventional collection selection techniques as complementary second-pass approaches to data-finding resources. Our longer-term view is to integrate such approaches into a suite of collection selection resources, both conventional and domain-driven. It would ultimately be up to users to determine which approach will be most appropriate for a given information need. Over time and with experience, they will best be able to judge, based on the granularity and context of the query, what would be the most reasonable approach (or utility) to invoke. In general, our methodology demonstrates how a user-centric approach can lead to long-term user satisfaction and search efficiency in a computationally inexpensive first-line-of-attack manner. The extent to which this approach is generalizable to nonprofessional domains remains an open research question.

The chief obstacles to developing a user query category-based system are the initial time required for query analysis and the domain expertise required for the design of the authority resources. In addition, managing updates in a rapidly growing data environment can also pose considerable challenges.

8. FUTURE WORK

As a result of our preliminary query analysis, we plan to add at least two additional query categories to our user resources, including *Financial* and *Publisher*, both of which arose directly from our query investigation work. Although *Financial* was a category included in our initial analysis, because it represented less than 2% of the total queries, it was not addressed in the preliminary system. Yet given large volumes of queries being submitted to a database selection system, even 1% of the total volume can represent, in absolute terms, a significant number of additional users being satisfied. We also plan on addressing issues related to query granularity and precision by creating regional contextual filters of existing authority resources. With this approach, users in the European Union, for example, will default to a different set of data authority resources than users in Australia and New Zealand, unless they indicate a preference for a broader nonregional view. This will require additional evaluations based on queries run against each individual region-specific data set. We intend to combine this approach with conventional collection selection and language modeling techniques to provide a seamless integrated array of information-finding tools. The particular utility a user selects would likely be decided by the nature of the query, for instance, whether it is abstract or concrete.

We are also examining the prospects for deriving user query categories automatically, using a clustering algorithm similar to what Xu et al. used when harnessing language modeling techniques for the collection selection problem [Xu and Croft 1999; Karypis 2002]. It remains unclear how satisfactory machine-generated classifications would be to users in specific professional domains. There may be room, however, for introducing a human in the loop to supervise the types of clusters that are being formed.

APPENDIX

A. INSTRUCTIONS TO DOMAIN EXPERTS FOR QUERY CATEGORY DETERMINATION TASK

You are being asked to develop categories for a few hundred sample user information requests. The queries come from the existing database selection system (IDEN). This should take you roughly a couple of hours. If it takes longer than that, you may be thinking too hard. Included below are some suggestions.

—To help categorize the request, you might first ask yourself, "What type of terms did the user enter?"

—If you are unable to categorize the type of terms, or if that doesn't provide enough detail, you might ask yourself, "What kind of information will provide an adequate response to this request?"

—Feel free to develop categories that go beyond document type. You could use a combination of the type of request and the suggested type of material. For example, if the user's search was for *42 USC 1395nn*, you could characterize that as "Citation to a specific document"—which could be more helpful than categorizing the request as simply "Statute" or "document citation."

—If you can't determine what a user is looking for, or even what type of information could answer the question (e.g., the term or concept is completely unknown), it is acceptable to conclude that you simply don't know.

—Our objective is to arrive at a reasonable number of categories, a set a user could look through and choose a category from without spending a lot of time to find the "correct" one.

—Given this decidedly finite number of sample requests, your list of categories may not be exhaustive. Do not be too concerned that after having worked through the list, you note that some significant category of information request is not represented. Please include it, given your knowledge about the domain and these types of requests.

APPENDIX

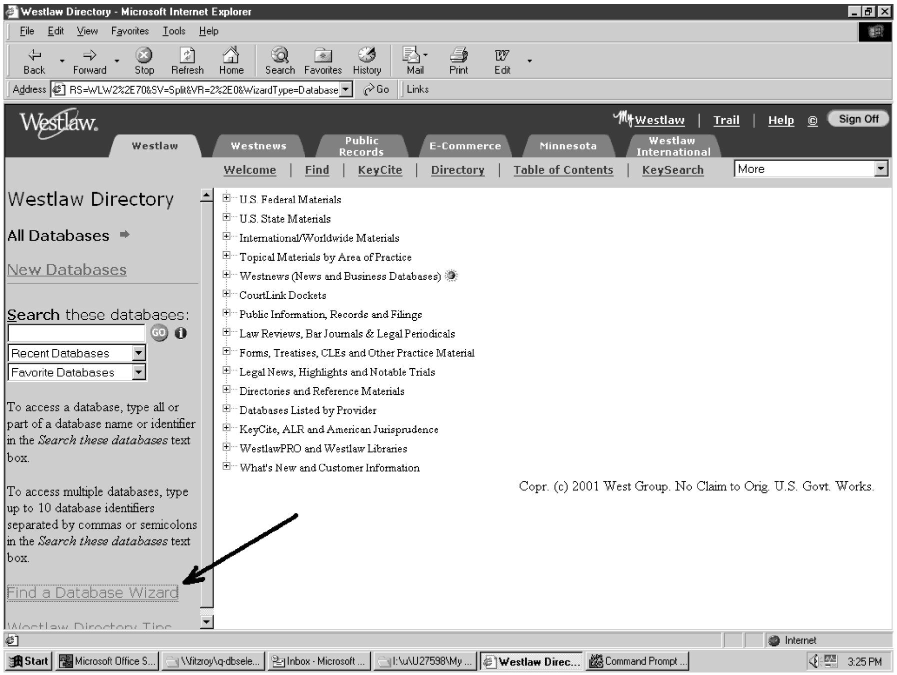
B. TYPICAL LEGAL PRACTICE AND RESEARCH AREAS

(1) Administrative Law
(2) Admiralty & Maritime Law
(3) Agriculture Law
(4) Alternative Dispute Resolution
(5) Antitrust & Trade Regulation
(6) Banking & Finance Law
(7) Bankruptcy Law
(8) Business & Commercial Law
(9) Business Organizations
(10) Civil Rights
(11) Communications & Media Law
(12) Constitutional Law
(13) Construction Law
(14) Criminal Law
(15) Cyberspace Law
(16) Education Law
(17) Election Campaign & Political Law
(18) Employment Law
  (a) Employee
  (b) Employer

(19) Energy Law
(20) Entertainment, Sports, & Leisure Law
(21) Environmental Law
(22) Estate Planning
(23) Ethics & Professional Responsibility
(24) Family Law
(25) Gaming Law
(26) Government Agencies & Programs
(27) Government Contracts
(28) Health, Medicine, & Health Care Law
(29) Immigration & Naturalization Law
(30) Insurance Law
(31) Intellectual Property Law
(32) International Law
(33) Labor Law
(34) Litigation & Appeals
(35) Military Law
(36) Native Peoples Law
(37) Natural Resources Law
(38) Personal Injury
    (a) Defense
    (b) Plaintiff
(39) Probate & Estate Law
(40) Products Liability Law
(41) Professional Malpractice Law
(42) Real Estate Law
(43) Science & Technology Law
(44) Securities Law
(45) State, Local, & Municipal Law
(46) Taxation Law
(47) Toxic Torts
(48) Transportation Law
(49) Workers' Compensation
(50) Year 2000 (Y2K) Law

APPENDIX

C. SAMPLE USER—SYSTEM INTERACTION FOR PRACTICE AREA QUERY


Westlaw Directory - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Forward Stop Refresh Home Search Favorites History Mail Print Edit
Address
Go
Links
Westlaw.
My Westlaw | Trail | Help ©
Sign Off
Westlaw
Westnews
Public Records
E-Commerce
Minnesota
Westlaw International
Welcome | Find | KeyCite | Directory | Table of Contents | KeySearch
More
GO
Internet
Start
Microsoft Office S...
\\fitzroy\q-dbsele...
Inbox - Microsoft ...
I:\u\U27598\My ...
Westlaw Direc...
Command Prompt ...
3:25 PM

Address: RS=WLW2%2E70&SV=Split&VR=2%2E0&WizardType=Database

Westlaw Directory

All Databases

New Databases

Search these databases:

Recent Databases

Favorite Databases

To access a database, type all or part of a database name or identifier in the *Search these databases* text box.

To access multiple databases, type up to 10 database identifiers separated by commas or semicolons in the *Search these databases* text box.

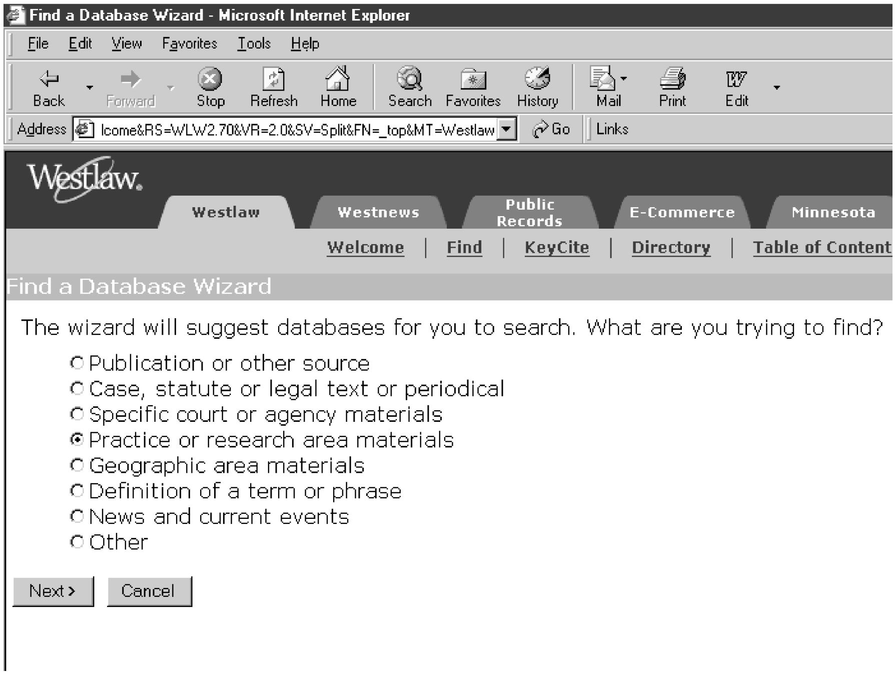
Find a Database Wizard

Westlaw Directory Tips

- U.S. Federal Materials
- U.S. State Materials
- International/Worldwide Materials
- Topical Materials by Area of Practice
- Westnews (News and Business Databases)
- CourtLink Dockets
- Public Information, Records and Filings
- Law Reviews, Bar Journals & Legal Periodicals
- Forms, Treatises, CLEs and Other Practice Material
- Legal News, Highlights and Notable Trials
- Directories and Reference Materials
- Databases Listed by Provider
- KeyCite, ALR and American Jurisprudence
- WestlawPRO and Westlaw Libraries
- What's New and Customer Information

Copr. (c) 2001 West Group. No Claim to Orig. U.S. Govt. Works.

Fig. C.1. Westlaw welcome screen and link to DBS prototype.


Find a Database Wizard - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Forward Stop Refresh Home Search Favorites History Mail Print Edit
Address
Go
Links
Westlaw.
Westlaw
Westnews
Public Records
E-Commerce
Minnesota
Welcome | Find | KeyCite | Directory | Table of Content

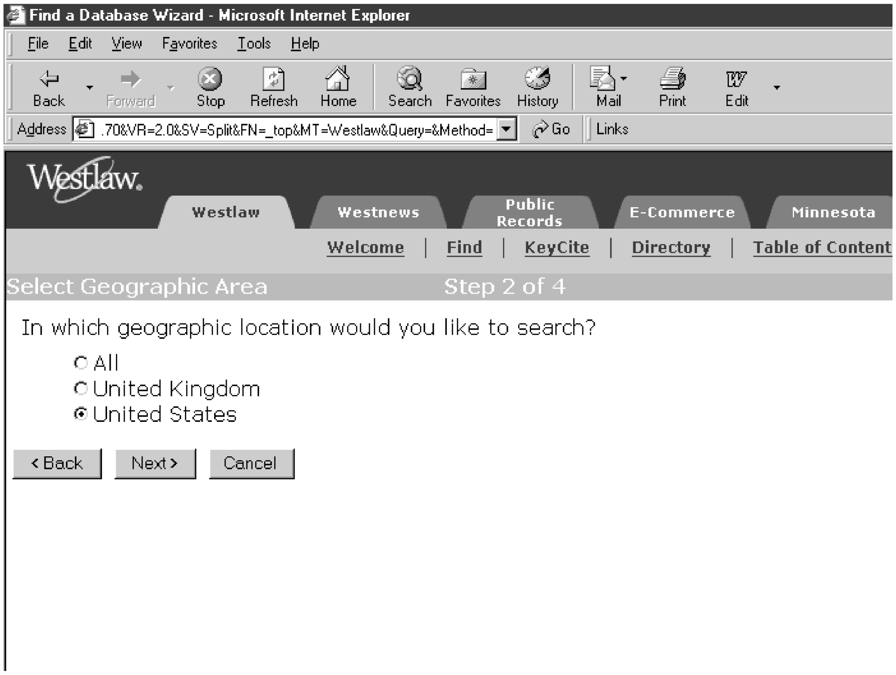
Address: lcome&RS=WLW2.70&VR=2.0&SV=Split&FN=_top&MT=Westlaw

Find a Database Wizard

The wizard will suggest databases for you to search. What are you trying to find?

- ○ Publication or other source
- ○ Case, statute or legal text or periodical
- ○ Specific court or agency materials
- ● Practice or research area materials
- ○ Geographic area materials
- ○ Definition of a term or phrase
- ○ News and current events
- ○ Other

Fig. C.2. DBS prototype, initial screen.


Find a Database Wizard - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Forward Stop Refresh Home Search Favorites History Mail Print Edit
Address
Go
Links
Westlaw.
Westlaw
Westnews
Public Records
E-Commerce
Minnesota
Welcome | Find | KeyCite | Directory | Table of Content

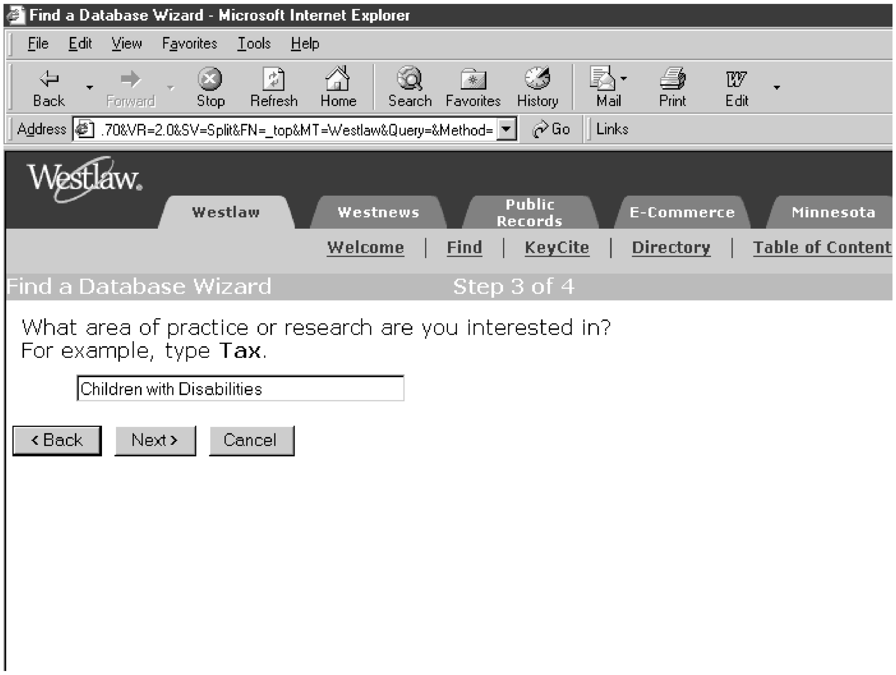
Address: .70&VR=2.0&SV=Split&FN=_top&MT=Westlaw&Query=&Method=

Select Geographic Area Step 2 of 4

In which geographic location would you like to search?

- ○ All
- ○ United Kingdom
- ● United States

Fig. C.3. DBS prototype, regional screen.

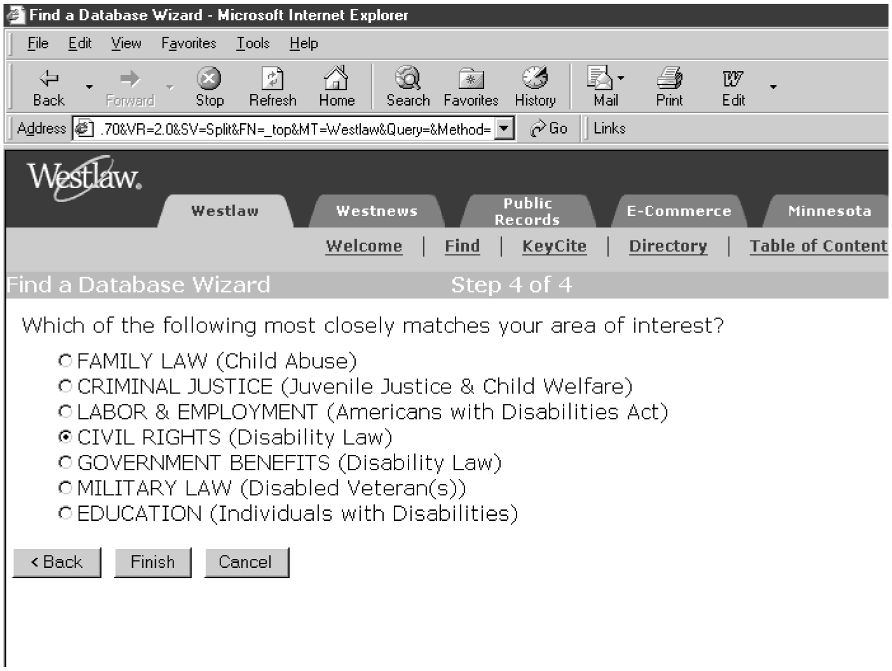

Find a Database Wizard - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Forward Stop Refresh Home Search Favorites History Mail Print Edit
Address .70&VR=2.0&SV=Split&FN=_top&MT=Westlaw&Query=&Method= Go Links
Westlaw.
Westlaw Westnews Public Records E-Commerce Minnesota
Welcome | Find | KeyCite | Directory | Table of Content
< Back Next > Cancel

Find a Database Wizard Step 3 of 4

What area of practice or research are you interested in?
For example, type **Tax**.

Children with Disabilities

Fig. C.4. DBS prototype, topical screen.

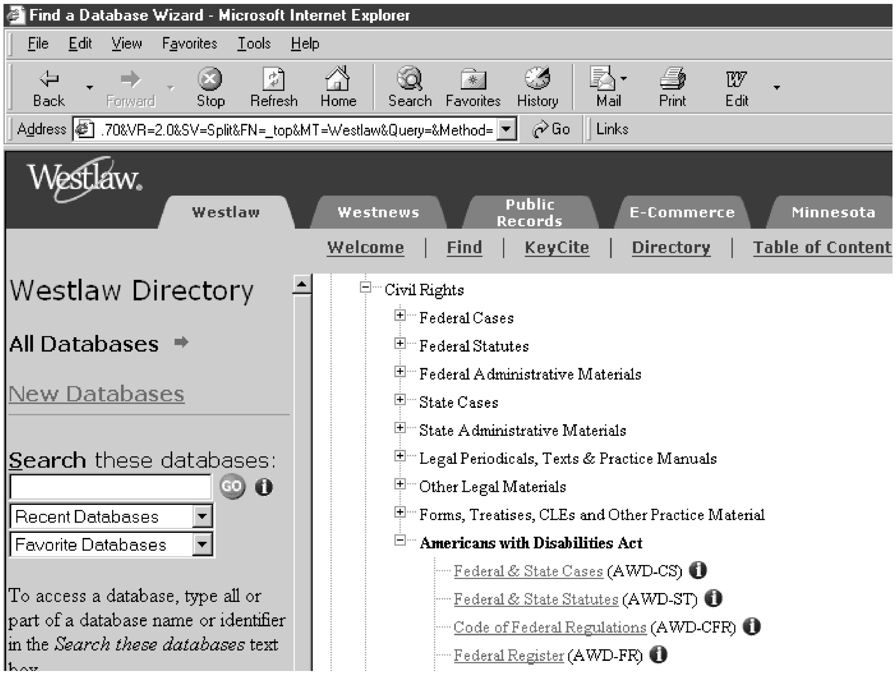

Find a Database Wizard - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Forward Stop Refresh Home Search Favorites History Mail Print Edit
Address .70&VR=2.0&SV=Split&FN=_top&MT=Westlaw&Query=&Method= Go Links
Westlaw.
Westlaw Westnews Public Records E-Commerce Minnesota
Welcome | Find | KeyCite | Directory | Table of Content
< Back Finish Cancel

Find a Database Wizard Step 4 of 4

Which of the following most closely matches your area of interest?

○ FAMILY LAW (Child Abuse)
○ CRIMINAL JUSTICE (Juvenile Justice & Child Welfare)
○ LABOR & EMPLOYMENT (Americans with Disabilities Act)
● CIVIL RIGHTS (Disability Law)
○ GOVERNMENT BENEFITS (Disability Law)
○ MILITARY LAW (Disabled Veteran(s))
○ EDUCATION (Individuals with Disabilities)

Fig. C.5. DBS prototype, topical selection screen.


Find a Database Wizard - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Forward Stop Refresh Home Search Favorites History Mail Print Edit
Address .70&VR=2.0&SV=Split&FN=_top&MT=Westlaw&Query=&Method=
Go
Links
Westlaw.
Westlaw
Westnews
Public Records
E-Commerce
Minnesota
Welcome | Find | KeyCite | Directory | Table of Content
Recent Databases
Favorite Databases

Westlaw Directory

All Databases

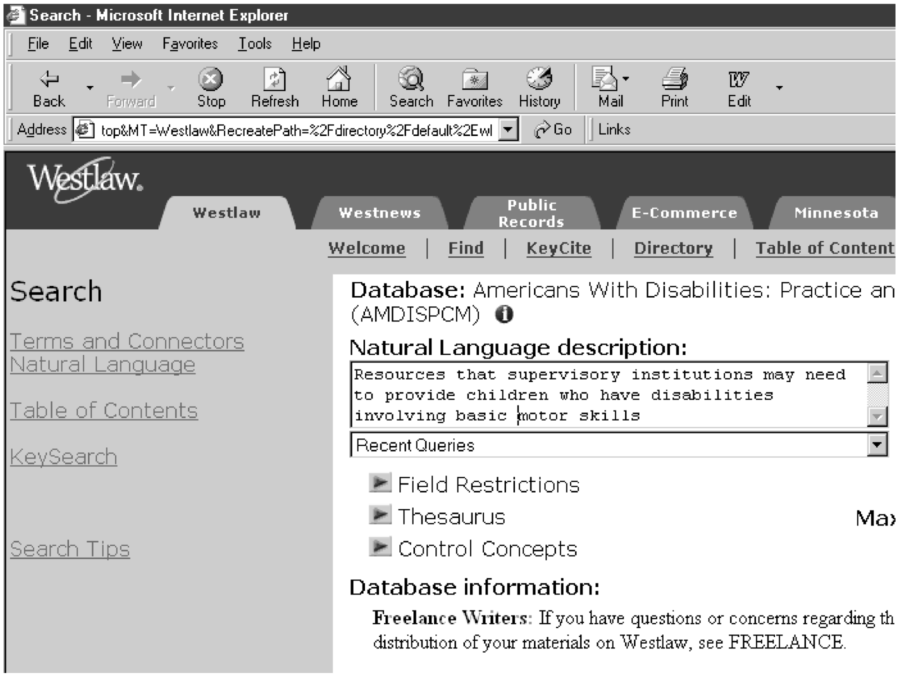
New Databases

Search these databases:

To access a database, type all or part of a database name or identifier in the *Search these databases* text box.

- Civil Rights
  - Federal Cases
  - Federal Statutes
  - Federal Administrative Materials
  - State Cases
  - State Administrative Materials
  - Legal Periodicals, Texts & Practice Manuals
  - Other Legal Materials
  - Forms, Treatises, CLEs and Other Practice Material
  - **Americans with Disabilities Act**
    - Federal & State Cases (AWD-CS)
    - Federal & State Statutes (AWD-ST)
    - Code of Federal Regulations (AWD-CFR)
    - Federal Register (AWD-FR)

Fig. C.6. DBS prototype, subdirectory screen for selected practice/research area.


Search - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Forward Stop Refresh Home Search Favorites History Mail Print Edit
Address top&MT=Westlaw&RecreatePath=%2Fdirectory%2Fdefault%2Ewl
Go
Links
Westlaw.
Westlaw
Westnews
Public Records
E-Commerce
Minnesota
Welcome | Find | KeyCite | Directory | Table of Content
Terms and Connectors
Natural Language
Table of Contents
KeySearch
Search Tips
Recent Queries
Field Restrictions
Thesaurus
Control Concepts
Ma

Search

Database: Americans With Disabilities: Practice an (AMDISPCM)

Natural Language description:

```
Resources that supervisory institutions may need
to provide children who have disabilities
involving basic motor skills
```

Database information:

Freelance Writers: If you have questions or concerns regarding th distribution of your materials on Westlaw, see FREELANCE.

Fig. C.7. DBS prototype, final user-selected database entry point.

ACKNOWLEDGMENTS

We thank Peter Jackson and Denis Hauptly for their ongoing support of this project. We also thank Howard Turtle for his initial promotion of both this research and our collaboration with groups providing user analysis. We appreciate the contributions of David Eckardt and Julie Gleason in helping identify our query categories. We are grateful to Bill Kahnk for his contribution of collection relevance judgments and his ongoing willingness to participate in interassessor validations. We also thank Pauline Afuso and Carol Cunningham for their participation in the interassessor validations. Thanks also to Dan Brick for his assessments of our statistical methods. To Jamie Callan, for his invaluable comments on this work, we are also thankful. Finally, we are grateful to the anonymous referees for their very useful and detailed comments on this article.

REFERENCES

ALLAN, J., CALLAN, J., CROFT, W. B., BALLESTEROS, L., BYRD, D., SWANN, R., AND XU, J. 1997. INQUERY does battle with TREC-6. In *Proceedings of the Sixth Text REtrieval Conference (TREC-6)*, NIST, 169–206.

BELKIN, N. J. 2000. Helping people find what they don't know. *Commun. ACM 43*, 8 (August), 58–61.

BOPP, R. E. AND SMITH, L. C. 2000. *Reference and Information Services, An Introduction* (3rd ed.). Libraries Unlimited, Englewood, CA, Chapter 3, Bibliographic Control, Organization of Information, and Search Strategies, 88–93.

BORLASE, R. 1999. KeyCite: Westlaw's new citator. University of Houston Law School. Available at http://www.law.uh.edu/guides/KeyCite.html, 1–2.

BROGLIO, J., CALLAN, J., AND CROFT, W. B. 1993. INQUERY system overview. In *Proceedings of the TIPSTER Text Program (Phase I)*, NIST, 47–67.

BUCKLAND, M., CHEN, A., CHEN, H.-M., KIM, Y., LAM, B., LARSON, R., NORGARD, B., AND PURAT, J. 1999. Mapping entry vocabulary to unfamiliar metadata vocabularies. *D-Lib Magazine*. Available at www.dlib.org.

BUCKLEY, C. AND VOORHEES, E. M. 2000. Evaluating evaluation measure stability. In *Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '00)* (Berkeley, CA), ACM, New York, 33–40.

CALLAN, J. AND CONNELL, M. 2001. Query-based sampling of text databases. In *ACM Trans. Inf. Syst. 19*, 2 (April), 97–130.

CALLAN, J., LU, Z., AND CROFT, W. B. 1995. Searching distributed collections with inference networks. In *Proceedings of the Eighteenth Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '95)* (Seattle, WA), ACM, New York, 21–29.

CARLETTA, J. 1996. Assessing agreement on classification tasks: The kappa statistic. *Comput. Ling. 22*, 2, 249–254.

CHAKRAVARTHY, A. S. AND HAASE, K. B. 1995. Netserf: Using semantic knowledge to find Internet information archives. In *Proceedings of the Eighteenth Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '95)* (Seattle, WA), ACM, New York, 4–11.

CONRAD, J. G. AND DABNEY, D. P. 2001. A cognitive approach to judical opinion structure: Applying domain expertise to component analysis. In *Proceedings of the Eighth International Conference of Artificial Intelligence and Law (ICAIL'01)* (St. Louis, MO), ACM, New York, 1–11.

CONRAD, J. G. AND UTT, M. H. 1994. A system for discovering relationships by feature extraction from text databases. In *Proceedings of the Seventeenth Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '94)* (Dublin), Springer-Verlag, New York, 260–270.

CONRAD, J. G., GUO, X. S., JACKSON, P., AND MEZIOU, M. 2002. Database selection using complete physical and acquired logical collection resources in a massive domain-specific operational environment. In *Proceedings of the 28th International Conference on Very Large Databases (VLDB'02)* (Hong Kong), Morgan-Kaufmann, San Francisco, 71–82.

DOLAN, R., AGRAWAL, D., DILLON, L., AND ABBADI, A. E. 1996. Pharos: A scalable distributed architecture for locating heterogeneous information sources. Tech. Rep. TRCS95-05, University of California–Santa Barbara, Department of Computer Science, July.

DOZIER, C. AND HASCHART, R. 2000. Automatic extraction and linking of person names in legal text. In *Proceedings of the RIAO* (*Computer Assisted Information Retrieval*) *Conference* (Paris), Content Based Multimedia Information Access, 1305–1321.

FAN, Y. AND GAUCH, S. 1999. Adaptive agents for information gathering from multiple, distributed information services. In *Proceedings of the American Association for Artificial Intelligence Spring Symposium in Intelligent Agents in Cyberspace (AAAI)* (Stanford University), AAAI Press, 40–46.

FIDEL, R. AND CRANDALL, M. 1998. The role of subject access in information filtering. In *Visualizing Subject Access for 21st Century Information Resources*, P. A. Cochran and E. H. Johnson, Eds., University of Illinois at Urbana-Champaign, 16–27.

FRENCH, J. C., POWELL, A. L., CALLAN, J., VILES, C. L., EMMITT, T., PREY, K. J., AND MOU, Y. 1999. Comparing the performance of database selection algorithms. In *Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '99)* (Berkeley, CA), ACM, New York, 238–245.

GAUCH, S., WANG, G., AND GOMEZ, M. 1996. ProFusion: Intelligent fusion from multiple, distributed search engines. *J. Univ. Comput. Sci. 2*, 9, 637–649.

GRAVANO, L., GARCIA-MOLINA, H., AND TOMASIC, A. 1994. The effectiveness of gloss for the text database discovery problem. In *Proceedings of the 1994 ACM SIGMOD International Conference on Management of Data (SIGMOD '94)* (Minneapolis), ACM, New York, 126–137.

HAWKING, D. AND THISTLEWAITE, P. 1999. Methods for information server selection. *ACM Trans. Inf. Syst. 17*, 1 (Jan.), 40–76.

HEARST, M. A. 1994. Using categories to provide context for full-text retrieval results. In *Proceedings of the RIAO (Computer Assisted Information Retrieval) Conference* (New York), CID, 115–130.

HEYDON, A. AND NAJORK, M. 1999. Mercator: A scalable, extensible Web crawler. *World Wide Web 2*, 4 (Dec.), 219–229.

KARYPIS, G. 2002. *CLUTO: A Software Package for Clustering High Dimentional Data*, Release 1.5 ed. Dept. of Computer Science, University of Minnesota, Minneapolis, Available at http://www-users.cs.umn.edu/~karypis/cluto.

KRAMER, J., NORONHA, S., AND VERGO, J. 2000. A user-centered design approach to personalization. *Commun. ACM 43*, 8 (August), 45–48.

LARKEY, L., CONNELL, M., AND CALLAN, J. 2000. Collection selection and results merging with topically organized U.S. patents and TREC data. In *Proceedings of the Ninth International Conference on Information Knowledge and Management (CIKM '00)* (McLean, VA), ACM, New York, 282–289.

MARCU, D. 2002. Personal communication. Information Sciences Institute (ISI), University of Southern California, Los Angeles.

MURRAY, B. H. AND MOORE, A. 2000. Sizing the Internet. A white paper, Cyveillance. July. Available at http://www.cyveillance.com/web/us/downloads/Sizing_the_Internet.pdf.

PARK, S. 2000. Usability, user preferences, effectiveness, and user behaviors when searching individual and integrated full-text databases: Implications for digital libraries. *J. Amer. Soc. Inf. Sci. 51*, 5 (March), 456–468.

PORTER, M. 1980. An algorithm for suffix stripping. *Program 14*, 3, 130–137.

POWELL, A. L., FRENCH, J. C., CALLAN, J., CONNELL, M., AND VILES, C. L. 2000. The impact of database selection on distributed searching. In *Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '00)* (Athens, Greece), ACM, New York, 232–239.

ROBERTSON, S. E., WALKER, S., JONES, S., HANCOCK-BEAULIEU, M. M., AND GATFORD, M. 1995. Okapi at TREC-3. In *Overview of the Third Text REtrieval Conference (TREC-3)* (Gaithersburg, MD), D. K. Harman, Ed., NIST, 109–126.

SANDERSON, M. 1994. Word sense disambiguation and information retrieval. In *Proceedings of the Seventeenth Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '94)* (Dublin), Springer-Verlag, New York, 142–151.

SARACEVIC, T. 1997. Users lost: Reflections on the past, present, future, and limits of information science. In *Proceedings of the Twentieth Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '97)* (Philadelphia), ACM, New York, 1–2.

SIEGEL, S. AND N. JOHN CASTELLAN, J. 1988. *Nonparametric Statistics for the Behavioral Sciences*. McGraw-Hill, Boston, Chapter 9: Measures of Association and Their Tests of Significance, 284–289.

STELLIN, S. 2000. E-commerce report: Internet companies learn how to personalize. *New York Times*, C8.

THOMPSON, P., TURTLE, H., YANG, B., AND FLOOD, J. 1995. TREC-3 ad hoc retrieval and routing experiments using the WIN system. In *Overview of the Third Text REtrieval Conference (TREC-3)* (Gaithersburg, MD), D. K. Harman, Ed. NIST, 211–217.

TURTLE, H. 1994. Natural language vs. Boolean query evaluation: A comparison of retrieval performance. In *Proceedings of the Seventeenth Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '94)* (Dublin), Springer-Verlag, New York, 212–221.

TURTLE, H. R. 1991. Inference networks for document retrieval. PhD Thesis, University of Massachusetts-Amherst.

VOORHEES, E. M. 1998. Variations in relevance judgments and the measurement of retrieval effectiveness. In *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '98)* (Melbourne, Australia), ACM, New York, 315–323.

VOORHEES, E. M. 2002. The philosophy of information retrieval. In *Proceedings of the Second Workshop of the Cross-Language Evaluation Forum (CLEF '01)* (Darmstadt, Germany), Springer, New York, 355–370.

VOORHEES, E. M. AND HARMAN, D. 2000. Overview of the Sixth Text REtrieval Conference (TREC-6). *Inf. Process. Manage. 36*, 1 (Jan.), 3–35.

WANG, W., MENG, W., AND YU, C. 2000. Concept hierarchy based text database categorization in a metasearch engine environment. In *Proceedings of the First International Conference on Web Information Systems Engineering (WISE '00)* (Hong Kong), Available at http://rakaposhi.eas.asu.edu/havasu.html.

WU, Z., MENG, W., YU, C., AND LI, Z. 2001. Towards a highly scalable and effective metasearch engine. In *Proceedings of the Tenth International World Wide Web Conference (WWW '01)* (Hong Kong), ACM, New York, 386–395.

XU, J. AND CALLAN, J. 1998. Effective retrieval with distributed collections. In *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '98)* (Melbourne, Australia), ACM, New York, 112–120.

XU, J. AND CROFT, W. B. 1999. Cluster-based language models for distributed retrieval. In *Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '99)* (Berkeley, CA), ACM, New York, 254–261.

Received December 2001; revised April 2002; accepted October 2002