Client–System Collaboration for Legal Corpus Selection in an Online Production Environment

Jack G. Conrad
Research & Development
Thomson Legal & Regulatory
St. Paul, Minnesota 55123 USA
Jack.G.Conrad@Thomson.com

Joanne R.S. Claussen
Westlaw Technology Development
Thomson–West
St. Paul, Minnesota 55123 USA
Joanne.Claussen@Thomson.com

Abstract

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 result sets, 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 interpreted and treated accurately 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.

Keywords

Database Selection, Query Categorization, User Interaction

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 the paper focus exclusively on the resulting prototype. A production implementation based upon our research and user acceptance of the production system are discussed later in the paper. The bedrock of our system is the WIN search engine[2] [28, 29], a close relative to the INQUERY engine developed at the Center for Intelligent Information Retrieval at the University of Massachusetts [1, 5]. 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.[3] Many of these techniques require extensive knowledge of the term and concept distribution in available collections either directly or through preliminary query-based sampling [25, 31, 9, 8]. Some of these techniques suggest that a reorganization of large amounts of data, by topical organization for instance, may improve overall retrieval performance [21]. 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, while collection selection is more specific (as in a collection of textual documents). In our case, use of 'database' can be misleading since 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 our collections are disjoint 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 paper, and aligned with the central thrust of this body of work, we use database selection and collection selection interchangeably as our primary research descriptors.

In order to address effectively the database selection problem in the legal domain, we have developed what is essentially a knowledge-based system. Given over 15,000 collections, we determined that the construction of, for example, a rule-based expert system, was beyond the scope of the challenge. We thus integrated an array of domain expertise into our prototype for the benefit of our users and in order to strengthen the focus upon specific, relevant, and useful legal corpora. In domains like law, where practitioners require "completeness" in response to queries (i.e., high recall *and* high precision), compromises in the name of optimiza-

[2] WIN stands for Westlaw Is Natural.

[3] In this paper, we will use collection to refer to a database of textual documents.

ICAIL-2003 Edinburgh, Scotland UK

tion, for instance, in terms of recall, are more difficult to justify.

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 extra-legal sources to supplement their legal research. Information vendors such as Thomson–West 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 specialized domains like law, with highly skilled professionals who are trained to be more selective about their search results, 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 10-15KB per page [22, 19]. With a rate of growth of over seven million new pages added per day, the Web was on track to double by mid-2001 [22]. These figures indicate that in 2001 there were in the range of 40 to 60 terabytes of indexable text on the Web. Thomson–West's alliance with Dialog puts their combined repositories at over 20 terabytes of data, corresponding to tens of thousands of databases. 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 non-trivial cost. Given users in the legal profession 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 [10]. 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 paper, we describe a database selection tool that leverages collection representations composed largely of metadata to address this selection problem.[4]

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 [26, 15]. User-centric groups as a whole have increased their focus on personalized and customized presentation of information access options[5] [20, 2]. 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 paper 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 the substantial analysis of real user queries that forms the foundation of all subsequent investigation. Section 4 describes our experimental methodology, including validation procedures. Section 5 briefly addresses our collection ranking algorithms and how they are distinguished from related approaches. Section 6 discusses our experiments and how we evaluated our approach in comparison with existing methods. Section 7 examines this technique in the context of complete user information-seeking sessions. Our conclusions and description of future work are presented in Sections 8 and 9.

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: (1) collection identification and/or representation; (2) query translation; (3) collection ranking; (4) collection selection; (5) searching the chosen collections; and (6) 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 [9, 16, 25]. This research leverages a considerable amount from fully automated 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, as we are optimizing the selection of distributed collections (servers in their case). [17]. Yet the majority of these works also acknowledge an untapped role for user interaction in the selection process.

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 [13, 18]. Park 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 [23]. 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 term frequencies in each collec-

[4] In another work, we describe an approach that relies on production-caliber collection-based language models [11].

[5] 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* [27].

tion. 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) [25, 21]. 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. 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." They totaled more than 8,000 queries and represented over 7,000 anonymous users. Approximately 7,500 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 1; see Figure 5 for examples). These designations represent important meta-level categories. We make no claim, however, to have identified or validated any sub-categories falling under these high-level designations. These categories include:

- document identifiers (e.g., by title or citation)
- named entities (e.g., person names or company names)
- 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)
- 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 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.

Our study 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 demonstrated that nearly 50% of our users' queries tended to mention a source, e.g., *Federal District Court Cases*, or a publication, e.g., *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, e.g., "Criminal Law" or "Alabama." For these types of queries, it might help if the user could be brought to some sort of central sub-directory 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, e.g., "Does an employee who slips on a wet floor have the right to proper compensation?" 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 [12, 14, 4]. In this user query analysis, a number of the 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,[6] as well as geographic regions/locations). Such metadata authority resources are discussed in more detail in the next section. 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.

No.	Category	Distribution
1.	Source or Publication ($\surd$)	48.2%
2.	Legal Issue ($\surd$)	13.2%
3.	Court or Gov't Agency ($\surd$)	7.5%
4.	Practice or Research Area ($\surd$)	7.3%
5.	Document by Citation (*)	4.5%
6.	Company Name (*)	4.5%
7.	Document by Title (*)	4.2%
8.	Definition ($\surd$)	3.6%
9.	Person Name (*)	3.0%
10.	Geographic Name ($\surd$)	2.8%
11.	News or Events ($\surd$)	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%

Table 1: Database Selection Queries by Frequency — where $\surd$ indicates use in final model (Section 4) and * indicates treatment by a parallel model (as shown on the bottom of Figure 1)

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 perform 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

[6]There are roughly 50 major 'practice' or 'research' areas that are referred to in the legal domain (e.g., bankruptcy, employment, malpractice).

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 managed to implement this model in a large production environment without the need for massive meta-searches and expensive meta-collection builds and updates.

4. EXPERIMENTAL METHODOLOGY

Our study has four phases. The *first phase* consists of the analysis and subsequent categorization of several thousand real user queries (Section 3). 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 4). 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 6.1).[7] 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. This analysis is presented below (Section 6.2).

The baseline system uses WIN's automatic selection and ranking of the top 20 collections and makes no differentiation between query types (Section 6). 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:[8]

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–System Operational Diagram). Half of our methods rely on 'search'; the other half rely on 'navigation,' by using attenuated decision trees (e.g., Figure 2)[9]. 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. One or more of these id-specified collections can be selected by the user and subsequently searched. For the other two authority resources, W_PRAC (legal practice areas) and W_GEOG (geographical locations), *directory links* to a topical or regional *sub-directory* 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., w.r.t. issues, definitions, news (e.g., Figure 2)]. In the decision tree mode, 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 & publications (15,042 profiles) and courts & government agencies (3,287 profiles), lists of top ranked collections would permit a user to directly submit queries to one or more relevant databases. For legal practice & research areas (1,352 profiles) and geographic locations & 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 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 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. For this reason, we propose interacting with the user earlier in the retrieval process, thus establishing a greater confidence in the suggested collections that merit further inspection.

[7] 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 section 4.1.

[8] 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 paper, as they are separate query-types.

[9] Figure 1: Regarding "Parallel Systems" on bottom of Flow Chart, when a user enters a bona fide document reference (e.g., *Roe v. Wade* or *142 Cal.App.2d 575*) or a person or company name, special recognizers flag the query and direct it to one of the parallel systems' entry points.

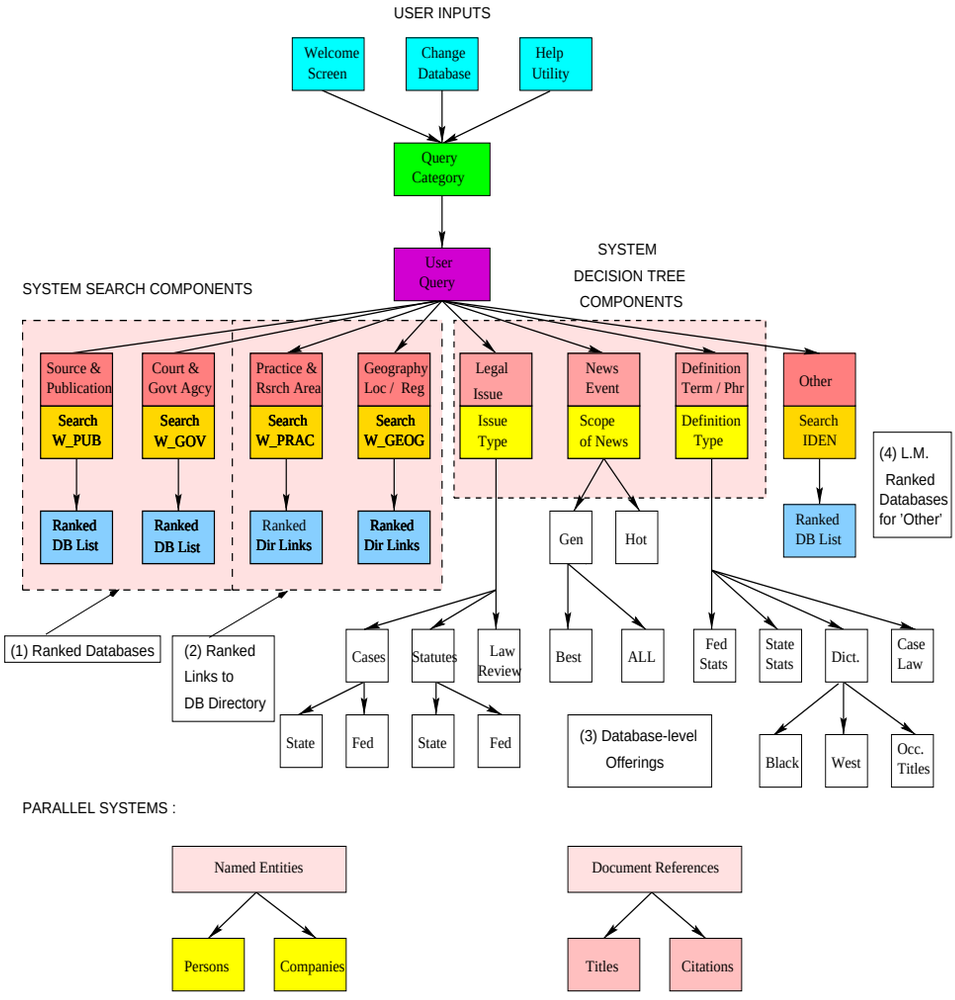

Figure 1: Flow Chart of Preliminary Operational System

4.1 Data

The metadata authority resources that support our searchable categories refer to to specialized sets of database profiles. Each corresponds to one of the meta-level categories discussed above. Whereas Buckland, et al. developed an Entry Vocabulary Technology to assist users in mapping their query vocabulary to that of potentially unfamiliar metadata vocabularies [6], 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 meta-level category that supports users' information needs. These include the following:

1. **W_PUB** (48.2%)[10]– 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 U.S. 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 (WL) 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 (WL) directory hierarchy.

[10]Figures in parentheses refer to percentage of overall DBS queries. [From Table 1]

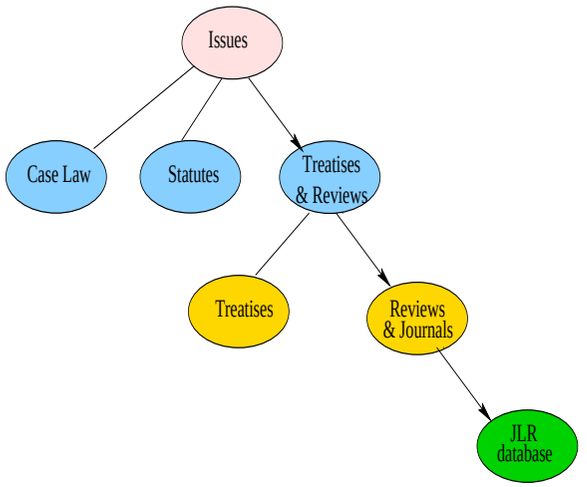

Figure 2: Sample Traversal for the Issues Category.

- **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.

4.2 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 2. IDEN, by contrast, is a general source identification data set and is comparable 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. This internal repository is not available for end-user searching. Human inspection of these fields occurs when concept supplementation is found to be useful.

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

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 Figure 3 and a W_GOV profile document in Figure 4. 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. Since 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 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 & W_GOV) and the *link-based* authority resources (W_PRAC & 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.

<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>

Figure 3: Facsimile International Review Collection Profile (for Publications)

<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>

Figure 4: Facsimile Environmental, Health, and Safty Regulations Profile (for Courts and Government Agencies)

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 agen-

Category	Data Set	Profiles	Size	Indexed Terms	Min./Max. Profile Length	Mean Length (Std. Dev.)
Source / Publications	W_PUB	15042	2.34MB	219817	11 - 6685	15 (94.7)
Court / Gov't Agencies	W_GOV	3287	733KB	91532	11 - 4747	28 (115.1)
Research / Practice Areas	W_PRAC	1352	139KB	11725	7 - 10	9 (2.6)
Geog. Regions / Locations	W_GEOG	292	60KB	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 2: Collection Statistics for Metadata Authority Resources

cies), 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., terrorism), associated profiles are removed or created in semi-automated 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 as that for the U.S.

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.

4.3 User Subscription

In many real user environments, it is reasonable to provide a two-step information retrieval process such as we propose here—the first to select databases, the second to perform a search. For users of very large proprietary systems like Thomson–West's and Dialog's, many thousands of online databases may be accessible, and as a result, users may be unfamiliar with all but a small portion of the available resources. Moreover, clients with different sized enterprises and different information needs commonly 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), while still others have basic coverage plans with the option to expand access on a per-need basis. Subscription arrangements and the consequent variability of cost are one reason why database selection can be a two-step process for users. Accordingly, clients play an important role in deciding on the scope of their research, based on perceived relevance, value and cost of accessible materials. Even though there may exist a background set of user profiles to assist in intelligently fielding queries, our current challenge has been to foster an expanded client-system *interaction*.

4.4 Indexing

For each of the authority resources, meta-level information is maintained in XML-like tag sets. To aid retrieval, the majority of these tagged elements are indexed though 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 meta-level profiles are shown in Figures 3 and 4. 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 title 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.[11]

4.4.1 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 (30-50KB vs. 10-15KB) [22, 19, 10]. Although there are circumstances in which we can and do use term distribution histograms to represent the vocabulary of a collection [11], 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, while 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

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

why it can be useful for clients to play a greater role in the steps that narrow their candidate collections when database selection is performed.

4.5 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 originate from a different month's database selection query log. Each of these logs contains queries that are assigned meta-level query categories corresponding to those in Table 1. 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 [7], 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. Further, 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 Tables 3 & 4). We run each of these sets against authority resources indexed for WIN-based retrieval. Samples of these queries can be seen in Figure 5. Queries not falling into our most frequently occurring categories, i.e., falling under the 'Other' category, are used in a standard collection selection test run which is beyond the scope of this paper. Average query lengths vary from 4-7 terms for W_PUB and W_GOV to 1-3 terms for W_PRAC and W_GEOG.

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

Figure 5: Sample Database Selection Queries by Category

4.6 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 4.1 and 4.2). The judgments are binary in nature (i.e., relevant/not relevant) and are performed by one attorney with a graduate degree in library science.

Because our users place a premium on precision at top ranks, we focus special attention on the top ranks in which relevant collections appear.[12] We report whether or not 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 documents 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. 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.

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 [30]. 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 relying on the contributions of legal database domain expertise. This approach was viewed as reasonable from both a practical and evaluative standpoint.

In addition, these 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 brief description of contents. These were judged for relevance in the same manner. This latter comparison was performed using the first set of 450 queries.[13]

5. 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 [8]. 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 [29, 5, 1]. 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 and 4). $tf \cdot idf$ scoring is applied to calculate the probability of relevance or belief score for a given collection profile, p_{bel}.

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

[12]The organization sponsoring our research also underscored the importance of tuning the system to produce superior top-rank precision, thus avoiding user frustration when obliged to examine unduly long lists.

[13]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.

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

$$idf_b = \frac{\log(\frac{N+0.5}{n})}{\log(N + 1.0)}$$

n represents the number of collection profile documents in which the query term, w_i, appears while 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 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 [24]. In addition, we rely on distilled consonant representations of terms present in the authority resources. The latter helps determine matches when users make common spelling errors or invoke non-standard abbreviations. We further apply query expansion techniques using a domain-specific thesaurus, acronym expansion, as well as other word forms when individuals use numerals or other terms with common or reasonable synonyms (e.g., *code* ⇔ *statutes*).

6. RESULTS

6.1 Individual Performance Evaluation

In the experiments described above, two sets of 450 category-specific queries (on sources, courts, practice areas, etc.) were run against the corresponding authority resources. Two separate months of real user queries were used for the task in order to determine whether there exist significant variances over time. These queries, harvested from query logs, were categorized by a legal practitioner commissioned to play the role of a representative user of the system. The results for the query sets are shown in Tables 3 and 4. 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.

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 between Tables 3 and 4).[14] The motivation for this non-standard approach to evaluation was four-fold. First, we did not have a fixed set of queries with mature, pre-existing relevance judgments, as in the case of some of the tracks at TREC conferences [30]. Second, we 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 evaluatory 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. Fourth, 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.

The classes are as follows:

1. — *single or all* relevant sources are present in ranks 1-5 (highest level precision); translates into 100% recall;

2. — *most* relevant sources in ranks 1-5, or single or all relevant sources in ranks 6-20 (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 1-5, 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 1-5, 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.

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 September's Court and Agency sets A & B, there do not appear to be any appreciable differences between each of the query categories' sets A & 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, 2, 3 and 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 is probably one reason why the system is usually able to capture the most salient database profiles in the top five ranks.

[14] The authors use a version of precision that is less granular than the standard definition. The two versions nonetheless share the same central notion of quantity relevant at rank n.

Category/ Test Set	Class							Total Queries
	1	2	3	4	4(a)	4(b)	4(c)	
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

Table 3: Performance Evaluation: Precision (September)

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/ Test Set	Class							Total Queries
	1	2	3	4	4(a)	4(b)	4(c)	
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

Table 4: Performance Evaluation: Precision (October)

6.2 Baseline Comparison

In addition to category-specific precision figures, we compare results from September's set of 450 queries with results obtained from the baseline IDEN system which also uses WIN ranking while relying upon one generic authority file of profiles. These results, shown in Table 5, 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) further 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 2). Hence, fewer non-relevant 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 W_GOV, has placed them among the top 1% most used collections of the more than 15,000 available on Westlaw.

7. DISCUSSION

This approach 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

Category/ Test Set	Class							Total Prcnt	Total Qrys
	1	2	3	4	4(a)	4(b)	4(c)		
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

Table 5: Performance Comparison: Category-based System vs. IDEN

are run against dedicated authority resources. Our studies show that these specific queries represent two-thirds of all queries. 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. Further, 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 "hidden" (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.[15]

Another issue we have had to address involves the overlap of our query categories (Table 1). 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 5). We are pursuing internal assistance to simulate such an evaluation.

8. CONCLUSIONS

We have shown that by permitting users to collaborate with an information finding system, users can find high-precision results effectively without the need for more computationally expensive mechanisms. Our approach is computationally inexpense 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 paper is novel in several respects.

- It is completely motivated by actual information needs expressed by users in the legal 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 themselves;
- This model attempts to bridge the divide that has long existed between computationally exhaustive systems deficient of any user-system interaction and information theories which stress the ongoing role of the user in search strategies.

Our results are consistent with Park's findings [23]—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

[15] 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 [3].

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 manner. The extent to which this approach is generalizable to non-professional 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.

9. ACKNOWLEDGEMENTS

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.

10. REFERENCES

[1] J. Allan, J. Callan, W. B. Croft, L. Ballesteros, D. Byrd, R. Swann, and J. Xu. INQUERY does battle with TREC-6. In *Proc. of the Sixth Text REtrieval Conf. (TREC-6)*, pages 169–206. NIST, Nov 1997.

[2] N. J. Belkin. Helping people find what they don't know. *ACM Communications*, 43(8):58–61, Aug 2000.

[3] R. E. Bopp and L. C. Smith. *Reference and Information Services, An Introduction (3rd ed.)*, chapter 3: Bibliographic Control, Organization of Information, and Search Strategies, pages 88–93. Libraries Unlimited, Englewood, CA, 2000.

[4] R. Borlase. KeyCite: Westlaw's new citator, University of Houston Law School. *Available at: www.law.uh.edu/ guides/KeyCite.html*, pages 1–2, 1999.

[5] J. Broglio, J. Callan, and W. B. Croft. INQUERY system overview. In *Proc. of the TIPSTER Text Program (Phase I)*, pages 47–67. NIST, 1993.

[6] M. Buckland, A. Chen, H.-M. Chen, Y. Kim, B. Lam, R. Larson, B. Norgard, and J. Purat. Mapping entry vocabulary to unfamiliar metadata vocabularies. *D-Lib Magazine*, www.dlib.org, Jan 1999.

[7] C. Buckley and E. M. Voorhees. Evaluating evaluation measure stability. In *Proc. of the 23rd Annual Int'l ACM SIGIR Conf. on Research and Dev. in IR*, pages 33–40. ACM Press, July 2000.

[8] J. Callan and M. Connell. Query-based sampling of text databases. *ACM Transactions on Information Systems (TOIS)*, 19(2):97–130, April 2001.

[9] J. Callan, Z. Lu, and W. B. Croft. Searching distributed collections with inference networks. In *Proc. of the 18th Annual Int'l ACM SIGIR Conf. on Research and Dev. in IR*, pages 21–29. ACM Press, July 1995.

[10] J. G. Conrad and D. P. Dabney. A cognitive approach to judical opinion structure: Applying domain expertise to component analysis. In *Proc. of the Eighth International Conf. of Artificial Intelligence and Law*, pages 1–11. ACM Press, May 2001.

[11] J. G. Conrad, X. S. Guo, P. Jackson, and M. Meziou. Database selection using complete physical and acquired logical collection resources in a massive domain-specific operational environment. In *Proc. of the 28th International Conf. on Very Large Databases*, pages 71–82. Morgan Kaufmann, Aug 2002.

[12] J. G. Conrad and M. H. Utt. A system for discovering relationships by feature extraction from text databases. In *Proc. of the 17th Annual Int'l ACM SIGIR Conf. on Research and Dev. in IR*, pages 260–270. Springer-Verlag, July 1994.

[13] R. Dolan, D. Agrawal, L. Dillon, and A. E. Abbadi. Pharos: A scalable distributed architecture for locating heterogeneous information sources. Tech Report TRCS95-05, University of California–Santa Barbara, Department of Computer Science, July 1996.

[14] C. Dozier and R. Haschart. Automatic extraction and linking of person names in legal text. In *Proceedings of the RIAO (Computer Assisted IR) Conference (Paris)*, pages 1305–1321. CID, April 2000.

[15] R. Fidel and M. Crandall. The role of subject access in information filtering. In P. A. Cochran and E. H. Johnson, editors, *Visualizing Subject Access for 21st Century Information Resources*, pages 16–27. University of Illinois at Urbana-Champaign, 1998.

[16] L. Gravano, H. Garcia-Molina, and A. Tomasic. The effectiveness of GlOSS for the text database discovery problem. In *Proc. of the 1994 ACM SIGMOD Int'l Conf. on Management of Data*, pages 126–137. ACM Press, May 1994.

[17] D. Hawking and P. Thistlewaite. Methods for information server selection. *ACM Transactions on Information Systems (TOIS)*, 17(1):40–76, Jan 1999.

[18] M. A. Hearst. Using categories to provide context for full-text retrieval results. In *Proc. of the RIAO (Computer Assisted IR) Conf. 1994*, pages 115–130. CID, Oct 1994.

[19] A. Heydon and M. Najork. Mercator: A scalable, extensible web crawler. *World Wide Web*, 2(4):219–229, Dec 1999.

[20] J. Kramer, S. Noronha, and J. Vergo. A user-centered design approach to personalization. *ACM Comm.*, 43(8):45–48, Aug 2000.

[21] L. Larkey, M. Connell, and J. Callan. Collection selection and results merging with topically organized U.S. patents and TREC data. In *Proc. of the 9th Int'l Conf. on Information Knowledge and Management*, pages 282–289. ACM Press, Nov 2000.

[22] B. H. Murray and A. Moore. Sizing the internet. A white paper, Cyveillance, July 2000.

[23] S. Park. Usability, user preferences, effectiveness, and user behaviors when searching individual and integrated full-text databases: Implications for digital libraries. *Journal of the American Society for Information Science*, 51(5):456–468, March 2000.

[24] M. Porter. An algorithm for suffix stripping. *Program*, 14(3):130–137, 1980.

[25] A. L. Powell, J. C. French, J. Callan, M. Connell, and C. L. Viles. The impact of database selection on distributed searching. In *Proc. of the 23rd Annual Int'l ACM SIGIR Conf. on Research and Dev. in IR*, pages 232–239. ACM Press, July 2000.

[26] T. Saracevic. Users lost: Reflections on the past, present, future, and limits of information science. In *Proc. of the 20th Annual Int'l ACM SIGIR Conf. on Research and Dev. in IR*, pages 1–2. ACM, July 1997.

[27] S. Stellin. E-commerce report: Internet companies learn how to personalize. *New York Times*, page C8, Aug 28, 2000.

[28] P. Thompson, H. Turtle, B. Yang, and J. Flood. TREC-3 ad hoc retrieval and routing experiments using the WIN system. In *Proc. of the Third Text REtrieval Conf. (TREC-3)*, pages 211–217. NIST, Nov 1995.

[29] H. R. Turtle. *Inference Networks for Document Retrieval*. Ph.D. dissertation., University of Massachusetts–Amherst, 1991.

[30] E. M. Voorhees and D. Harman. Overview of the Sixth Text REtrieval Conference (TREC-6). *Information Processing and Management*, 36(1):3–35, Jan 2000.

[31] J. Xu and J. Callan. Effective retrieval with distributed collections. In *Proc. of the 20th Annual Int'l ACM SIGIR Conf. on Research and Dev. in IR*, pages 112–120. ACM Press, Aug 1998.