Validation: A Critical First Step in the Evaluation of Systems for Legal Corpus Determination

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, both proprietary and Web-based, increases the importance of effective and efficient legal corpus selection and searching. Current "database selection" 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 data sets with relevant documents are not searched (compromised recall). It also merges result sets, often from disparate data sources, some that users may have discarded before their source selection task completed (diluted precision). We examine the impact that user interaction can have on the process of legal corpus 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. As a precursor to evaluation, in this workshop, we present three behind-the-scenes system validation exercises to assist us in determining whether certain system design decisions are justified in the context of our long-term goals of providing a corpus selection tool to legal practitioners. We ultimately show that by avoiding a one-size-fits-all approach that restricts the role users can play in information discovery, legal corpus selection effectiveness can be appreciably improved.

Keywords

Legal Corpus Selection, Query Categorization, User Interaction, Validation and Evaluation

1. INTRODUCTION

We have developed a model for improved legal corpus selection that offers the practitioner a key role in the discovery process. The model is based on the recognition that legal research 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 work serve to inform and validate design decisions for the resulting prototype. A production implementation based upon our research and user acceptance of the production system is discussed in more detail in our conference paper [8].

Designing and developing a comprehensive, reliable and practical legal research tool often poses distinct challenges, especially in terms of validating its usefulness, verifying its results, and maintaining its data integrity, especially for those systems that are knowledge-based. Moreover, the final evaluation of a system is not necessarily meaningful if the underlying assumptions that impact its design and development are not valid—and if human practitioners cannot use the results of these assumptions, presumably incorporated into the system, coherently and consistently.

Our prototype depends extensively upon certain canonical legal research tasks that are performed either by users themselves or by domain experts who are employed to evaluate our model, for instance, to judge experimental result sets for their precision and recall. The underlying notion that serves to motivate the three sets of validation exercises reported on here is that if we cannot demonstrate that legal practitioners themselves (i.e., representative system users) can reliably and repeatedly perform certain essential tasks, such as:

1. granularity assessments;
2. categorization;
3. relevance judgments;

then designers and evaluators cannot realistically come to expect computer systems to deliver consistent and satisfactory performance to their users.

Our resultant system relies upon the WIN search engine[2] [20, 21], a close relative to the INQUERY engine developed at the Center for Intelligent Information Retrieval at the University of Massachusetts [1, 3]. 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 corpus 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 [15, 25, 6, 5]. 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 [26, 12]. 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. Subsequent exploration of these issues has required us to first probe and validate some underlying assumptions any resultant system might depend upon.

In order to effectively address the challenge of corpus selection in the legal domain, we have developed what is essentially a knowledge-based system. Required to handle over

[2]WIN stands for Westlaw Is Natural.

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

15,000 collections, we determined that the construction of a rule-based expert system was beyond the scope of the problem. We were thus interested in marshalling domain expertise into a 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 optimization in terms of recall, for instance, are more difficult to justify. For this reason, in developing a reliable system for legal practitioners, we were required to be sensitive to both precision and recall.

Another significant aspect of the model involves the contribution of users. The retrieval community has repeatedly called for an increased role for users in Information Retrieval (IR) systems that are more effective than either computer-centric or user-centric approaches alone [17, 10]. User-centric groups as a whole have increased their focus on personalized and customized presentation of information access options[4] [11, 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 6 to 12 high-level classes of queries. Users have subsequently found this approach extremely useful.

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 expand beyond 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 typically unacceptable.

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. In this paper, we describe a corpus selection tool that leverages collection representations composed largely of metadata to address this selection problem.[5]

The remainder of this work is organized as follows: Section 2.1 addresses our exercise with domain experts to determine both an appropriate number and granularity of query categories in the system. Section 2.2 explores the degree to which legal practitioners can be expected to concur in their query categorization task. And finally, Section 2.3 examines concordances between domain experts who have been commissioned to provide relevance judgments for databases suggested by our system during its evaluation phase. Our summary remarks are provided in Section 3 and our discussion of long-term goals in Section 4.

[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* [19].

[5]In another work, we describe an approach that relies on production-caliber collection-based language models [9].

2. EXPERIMENTAL METHODOLOGY

The study we report on in our conference paper consists of four phases [8]. The *first phase* consists of the analysis and *validation* of legal research categories and the categorization of several thousand real user queries (Table 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 2.1.1). In the *third phase*, we enlist two sets of 450 real user queries that meet certain query category criteria and run those queries against metadata authority resources (databases) derived from the previous phase.[6] This phase also includes two additional *validation* steps involving real user queries, domain expert input, and correlation measures to test the reliability of the model's underlying assumptions (Sections 2.2.1-2 and Section 2.3.1-2). 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 [8].

2.1 User Query Analysis

Approximately two weeks of real users' database selection descriptions were initially inspected. Users submitted them to a system by selecting a button labeled "Search for a Database." The queries totaled more than 8,000 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 1 for examples). These designations represent important meta-level categories.[7]

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)

[6]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.

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

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 2.1.1) and * indicates treatment by a parallel model (as illustrated in Figure 2 in Appendix B)

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, etc). 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.

2.1.1 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[8] or as a query log analyst. In this exercise, we provided the three with a diverse set of 200 corpus selection queries. They came from a larger set of real user requests randomly selected from a query log associated with an existing corpus 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 2.[9]

[8]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.

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

From the findings presented in Table 2, we see that domain expert #1 suggests the least fine-grained categories while domain expert #3 contributes the most fine-grained, with domain expert #2's contributions representing something in between. Further, 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*). Since 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. The resulting primary source of categories is represented in Table 1.

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 1: Sample Database Selection Queries by Category

2.2 System 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 corpus 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 corpus selection application known as IDEN (short for Identification). 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 [4], 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 source & publication query sets are 50% larger than those for the other categories since our analysis showed that nearly 50% of all corpus selection queries appeared under this classification. The categories we use in these experiments include (1) sources/publications, (2) courts/government agencies,

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 ⋆ foreign country
⋆ specific database identifiers		⋆ a database name or id
	⋆ definitions	⋆ a noun
	⋆ news & events	
	⋆ financial queries	⋆ statistics
⋆ indeterminable	⋆ category unclear	⋆ unknown

Table 2: Corpus Selection Query Categories – Domain Expert Responses

(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). We run each of these sets against authority resources indexed for WIN-based retrieval. Samples of these queries can be seen in Figure 1. 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 sources & publications (W_PUB) and courts & government agencies (W_GOV) to 1-3 terms for legal research & practice areas (W_PRAC) and geographical locations & regions (W_GEOG).

2.2.1 *Validation of Query Categorization*

Because these queries were initially categorized by an attorney who is familiar with our meta-level 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

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

Table 3: Query Categorization: Expertise among Legal Practitioner–Participants, where D.B.S.= Database Selection; L.S.=Library Science.

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

Table 4: Kappa Statistics for Categorization Performed by Four Assessors [$z = 2.32$ **for** $\alpha = 0.01$] **.** $\kappa = 1$ **for complete agreement among assessors;** $\kappa = 0$ **for no agreement among assessors.**

paralegal and three attorneys. Of the attorneys, two had no familiarity with corpus 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 that use a system like Westlaw (Table 3). In this experiment, each of the participants was given a set of 200 real user queries and asked to categorize them using the first twelve categories shown in Table 1. 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 the number of tokens = 1, 2, 3, 4, and 5 or more tokens). Only one pretrial inspection of the collected user queries that was made in order to determine whether or not the query would be reasonably interpretable by someone with some degree of legal training. Hence a query like "alsdkjf" would be discarded.

To compare our inter-assessor concordances for the 200 queries and the 12 categories, we used the Kappa statistic for nominally scaled data [18]. We explored inter-assessor agreement relative to query length since it has been shown that longer query statements reduce the ambiguity associated with very short queries [16]. 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 4.

Computational Linguists have taken $\kappa = 0.8$ as the norm for significantly good agreement, though some argue that there is insufficient evidence to choose 0.8 over, for instance, other values between 0.6 and 0.9 [13]. To underscore the significance of these 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).[10] 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.

[10]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.

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, there also appears to be a relationship between query length and query clarity or disambiguation [16].

2.2.2 *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 which 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 $H_1 : \kappa > 0$ [7, 18].

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 4 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 corpus selection search.

2.3 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 (e.g., those illustrated in Figure 1). 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. We report whether or not relevant results are in the top 5 as well as the top 20 ranks and

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 Research & Practice Area	8/10 (80%)	133/135 (98.5%)
Geographical Location & Region	10/10 (100%)	10/10 (100.0%)
Combined	34/50 (68%)	607/663 (91.6%)

Table 5: Relevance Judgment Concordances

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

Table 6: Significance Tests for Relevance Judgment Concordances

if those results include one, some, all, or none of the relevant collections available. We place this emphasis on the top five collections because users looking for relevant corpora with which to begin their research have little patience to examine 19 of 20 candidates before encountering the first corpus containing relevant documents. This top 20 analysis and evaluation is performed on a total of 900 queries (i.e., two groups of 450 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.

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 [24]. Because of the nature of our relationship with the sponsoring department, a production-oriented business unit, 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 [23], 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 [8]. These were judged for relevance in the same manner. This latter comparison was performed using the first set of 450 queries.[11]

2.3.1 *Inter-Assessor Validation*

In order to perform a preliminary investigation into the dependability of the relevance judgments, a second attorney with no background in library science participated in an inter-assessor 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 corpora returned in response to 50 actual user queries. The queries included four sets of ten queries, one set for each authority resource, plus an additional ten from the most dominant category, sources & publications. All were randomly selected from their query category. Judgments were made on the top 20 corpora returned. For some queries, less than 20 corpora were returned, especially in the case of geographical locations where fewer possible matches can occur. In this inter-assessor correlation study, the two attorneys assessed the results and were in agreement for 92% of the queries (Table 5).

The second column in Table 5 represents the percentage of queries for which the assessors were in complete agreement.

2.3.2 *Testing the Significance of the Inter-Assessor Concordances*

The null hypothesis for these inter-assessor 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 in so far 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 6). By contrast,

[11] 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.

the Wilcoxon test favors the alternative hypothesis (one of the assessors will produce more positive relevance judgments than the other) since 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 5). 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 U.S.-based, and non-U.S.-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. 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" [22]. 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.

3. SUMMARY

In this work and in our longer conference paper, we show that by permitting users to collaborate with an information finding system, users can experience higher precision results without the need for more computationally expensive mechanisms. Our approach is computationally inexpensive insofar as it relies on authority resources that consist of data sets containing concise collection profiles. Such a well-conceived and experimentally validated approach 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 our work is novel in several respects.

- It is completely motivated by actual information needs expressed by users in the legal domain;
- In at least three important 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 [14]—that more user input is important in large environments with distributed data, that distinct corpus characterization can assist users with choices for integration, and that certain users prefer more *interactive* control over their selection processes. These findings also call into question the assumption that interaction with merged data is most effective (which is analogous to how prominent search engines field their results). Our model and resultant prototype is nonetheless bolstered by the series of critical validation exercises performed; they provide added credibility to the evaluation procedures that follow.

4. FUTURE WORK

Our long term goal is to integrate such approaches into a suite of collection selection resources, both conventional and domain-driven. It would ultimately be left 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 query context, what would be the most reasonable approach (or tool) to harness. 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.

5. 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. We are grateful to Pauline Afuso, Carol Cunningham and Bill Kahnk for their participation in our inter-assessor validations. We also appreciate the assessments of our statistical methods that Dan Brick provided.

6. 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 *Proceedings of the Sixth Text REtrieval Conference (TREC-6)*, pages 169–206. NIST, November 1997.

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

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

[4] C. Buckley and E. M. Voorhees. 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)*, pages 33–40. ACM Press, July 2000.

[5] J. Callan and M. Connell. Query-based sampling of text databases. In *ACM Transactions on Information Systems (TOIS)*, pages 97–130. ACM Press, April 2001.

[6] J. Callan, Z. Lu, and W. B. Croft. Searching distributed collections with inference networks. In

Proceedings of the 18th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '95) (Seattle, WA), pages 21–29. ACM Press, July 1995.

[7] J. Carletta. Assessing agreement on classification tasks: The kappa statistic. *Computational Linguistics*, 22(2):249–254, 1996.

[8] J. G. Conrad and J. R. Claussen. Client–system collaboration for legal corpus selection in an online production environment. In *Proceedings of the Ninth International Conference of Artificial Intelligence and Law (ICAIL'03) (Edinburgh, Scotland)*. 10 pages, ACM Press, June 2003.

[9] 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 *Proceedings of the 28th International Conference on Very Large Databases (VLDB'02) (Hong Kong)*, pages 71–82. Morgan Kaufmann, August 2002.

[10] 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.

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

[12] L. Larkey, M. Connell, and J. Callan. Collection selection and results merging with topically organized U.S. patents and TREC data. In *Proceedings of the 9th International Conference on Information Knowledge and Management (CIKM '00) (McLean, VA)*, pages 282–289. ACM Press, November 2000.

[13] D. Marcu. *Personal Communication*. Information Sciences Institute (ISI), University of Southern California, Los Angeles, CA, April 2002.

[14] 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.

[15] A. L. Powell, J. C. French, J. Callan, M. Connell, and C. L. Viles. 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)*, pages 232–239. ACM Press, July 2000.

[16] M. Sanderson. Word sense disambiguation and information retrieval. In *Proceedings of the 17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '94) (Dublin, Ireland)*, pages 142–151. Springer-Verlag, July 1994.

[17] T. Saracevic. Users lost: Reflections on the past, present, future, and limits of information science. In *Proceedings of the 20th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '97) (Philadelphia, PA)*, pages 1–2. ACM Press, July 1997.

[18] S. Siegel and J. N. John Castellan. *Nonparametric Statistics for the Behavioral Sciences*, chapter 9: Measures of Association and Their Tests of Significance, pages 284–289. McGraw Hill, Boston, 1988.

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

[20] P. Thompson, H. Turtle, B. Yang, and J. Flood. TREC-3 ad hoc retrieval and routing experiments using the WIN system. In D. K. Harman, editor, *Overview of the Third Text REtrieval Conference (TREC-3)*, pages 211–217. NIST (Gaithersburg, MD), April 1995.

[21] H. R. Turtle. *Inference Networks for Document Retrieval*. PhD thesis, University of Massachusetts-Amherst, 1991.

[22] E. M. Voorhees. 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)*, pages 315–323. ACM Press, August 1998.

[23] E. M. Voorhees. The philosophy of information retrieval. In *Proceedings of the Second Workshop of the Cross-Language Evaluation Forum (CLEF '01) (Darmstadt, Germany)*, pages 355–370. Springer, September 2002.

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

[25] J. Xu and J. Callan. 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)*, pages 112–120. ACM Press, August 1998.

[26] J. Xu and W. B. Croft. 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)*, pages 254–261. ACM Press, August 1999.

7. APPENDICES

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" – that 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

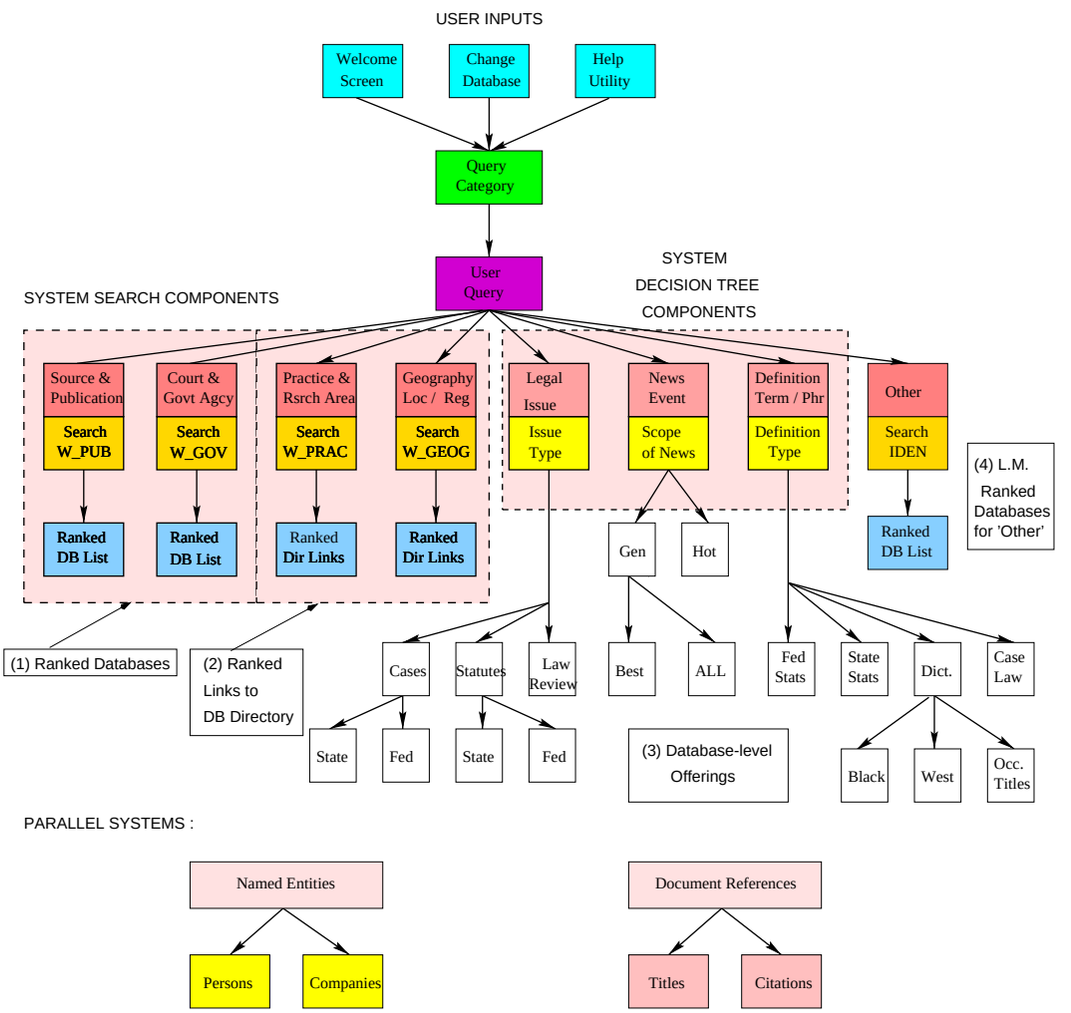

Figure 2: Flow Chart of Preliminary Operational System