Effective Collection Metasearch in a Hierarchical Environment: Global vs. Localized Retrieval Performance

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

Changwen Yang
Oracle Corporation
500 Oracle Parkway
Redwd Shores, CA 94065
Changwen.Yang@Oracle.com

Joanne S. Claussen
Technology Product Dev.
West Group
Eagan, MN 55123 USA
Joanne.Claussen@WestGroup.com

ABSTRACT

We compare standard global IR searching with user-centric localized techniques to address the *database selection problem*. We conduct a series of experiments to compare the retrieval effectiveness of three separate search modes applied to a hierarchically structured data environment of textual database representations. The data environment is represented as a tree-like directory containing over 15,000 unique databases and over 100,000 total leaf nodes. Our search modes consist of varying degrees of *browse and search*, from a global search at the root node to a refined search at a sub-node using dynamically-calculated inverse document frequencies (*idfs*) to score candidate databases for probable relevance. Our findings indicate that a browse and search approach that relies upon localized searching from sub-nodes is capable of producing the most effective results.

Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval—*search process,selection process*

General Terms

Performance, Experimentation, Human Factors

1. INTRODUCTION

The continued growth of online databases has made the work of finding the most relevant collections increasingly difficult. Until recently, the ability to execute a 'search' in a database directory as well as 'drill down' into its hierarchical structure have largely been regarded as separate activities. If either approach does not provide desired results, large numbers of users exit online systems with unmet information needs. *Yahoo!* and the *Open Directory Project* are exceptions that permit integrated browse and search. Research has begun to explore categorization and retrieval in such environments [4]. We hypothesized that if users could first browse to a potentially relevant sub-node in a large directory, results from a search in the sub-directory would be more precise than results from a search in the entire directory. To test the effectiveness of browse plus search functionality, we designed and conducted a series of experiments on three search modes. Using the same set of real user queries, these search modes included: (1) a global search of the directory from the root node, (2) a localized search of the relevant sub-directories using global idfs, and (3) a localized search of the relevant sub-directories using the appropriate dynamically-calculated local idfs.

2. OPERATIONAL ENVIRONMENT

The *Westlaw* system actually provides both a dedicated (single database search) and integrated (multiple database search) environment, while delivering merged document result lists. Park found that users ultimately desire more control over their searches where databases are concerned [2]. We wanted to pursue this issue of flexibility in alternative ways. To expand upon integrated versus dedicated database search capability and notions of simplicity plus control, we wanted to investigate another type of hybrid system, one which offers users effective browse *and* search functionality. We first wanted to inspect whether results from suitably restricted searching (following a user's navigation into a hierarchically arranged directory structure) would be measurably better than those from a simple global search at the root level. Secondly, we wanted to determine whether restricted search using standard collection-wide idf values or using localized idf values (defined by the sub-collection being searched) provided better performance. In both of these investigations, we wanted to determine whether higher precision could be achieved without a substantial loss of recall.

2.1 Data

West Group's data repository now contains over 5 million published case law opinions from virtually every state and federal jurisdiction. In the past decade, however, the amount of *non-legal* data available on Westnews, Westlaw's news-based data, has been growing at nearly exponential rates and now consists of millions more documents. *The Westlaw Directory* organizes over 15,000 unique databases in a tree-like structure and lists roughly 100,000 databases at its leaf nodes (with an average of 20,000 documents per database). In some regions, the Directory is 12 or more levels deep, but it averages 4-6 levels. Overall, it represents several terabytes of data. Available for use in this project are database profiles consisting of metadata documents containing the titles, topical coverage, and additional significant content associated with each database. Profile lengths vary from tens of core terms to several thousand characteristic tokens, depending on a database-related set of criteria.

#	User Queries	#	User Queries
1.	Labor Relations	5.	Compensation Planning
2.	State Legis. News	6.	Board of Immigration
3.	Voting Rights	7.	Food & Drug Admin.
4.	Admin. Materials Relating to Specific Occupations		

Table 1: ***Sample Queries***

2.2 Queries

We began with a set of 50 real user queries from an existing database selection system's query logs. We examined

SIGIR'02, August 11-15, 2002, Tampere, Finland.
ACM 1-58113-561-0/02/0008.

the granularity of the queries and determined that roughly 60% of them focused on mid-level nodes of the Directory.[1] This task was performed by a legal domain expert with a background in library science. The average query length was over three terms before stop words were removed and slightly under three terms after (Table 1).

3. SYSTEM ENVIRONMENT

Our system harnessed the WIN search engine[2], a cousin to the INQUERY engine developed at UMass–Amherst. INQUERY's and WIN's algorithms for ranking documents have been previously reported [1, 3], although we modified WIN's scoring formula for our last search mode which incorporated the use of dynamically calculated sub-directory specific local idf values (computed only when necessary for the user-selected sub-directory). WIN was run against all or a portion of the Westlaw Directory data (i.e., the database profiles described in section 2.1 and subsequently judged for relevance by the domain expert). Below is the $tf \cdot idf$ expr. used.

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

$$tf_b = d_t + (1-d_t)\cdot \frac{\log(tf_i+0.5)}{\log(tf_{max}+1.0)} \quad and \quad \underline{idf_b} = \frac{\log(\frac{N+0.5}{n})}{\log(N+1.0)}$$

d_b is the minimum belief component and d_t is the minimum term frequency component when term w_i is present in a database profile document, d_j. The value tf_{max} is the frequency of the most frequently occurring term in the corpus. n represents the number of database profile documents in which the query term w_i appears while N is the total number of database profile 'documents.'

4. EXPERIMENTS

We ran the set of resultant queries in three search modes: (1) global search of the entire data hierarchy using global idfs; (2) local search of the relevant sub-directories using global idfs; (3) local search of the relevant sub-directories using the appropriate dynamically calculated local idfs; where the $\underline{idf_b}$ parameter (above) was modified depending on the currently defined scope of the profile set, either global or local.

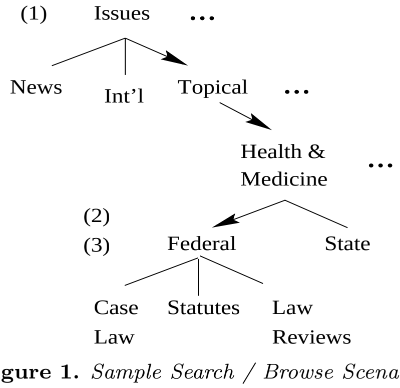

Figure 1. *Sample Search / Browse Scenario*

The "relevant sub-directories" were determined by a domain expert assigned to the project. They were selected as the first sub-directory or sub-directories under which a user could expect to find the initial relevant materials [e.g., (2) and (3) in Figure 1]. Emphasis was placed on precision at 20 and recall at 20 databases returned, since there is evidence that users generally do not examine candidate databases bevond this initial set.

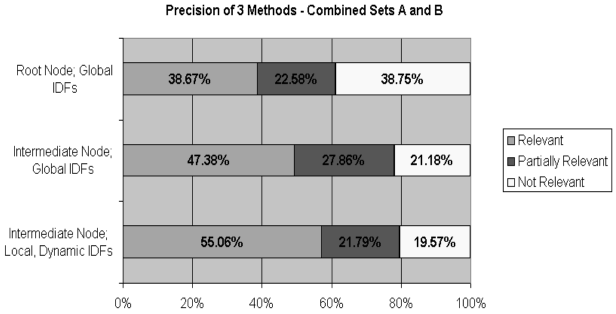

Figure 2. *Search Mode Results (Precision)*[1]

5. RESULTS

Our experimental findings suggest that browse+refined search can improve precision in relation to conventional global searches in environments where the data is organized in a hierarchical fashion such as in the Westlaw Directory. We further determined that restricted searches using localized idfs defined by the applicable sub-branches perform better than restricted searches using the more easily obtainable global idfs (Figure 2, with improvements of as much as 40%) with little evidence of a meaningful loss in recall (Figure 3, with losses averaging around 15%).

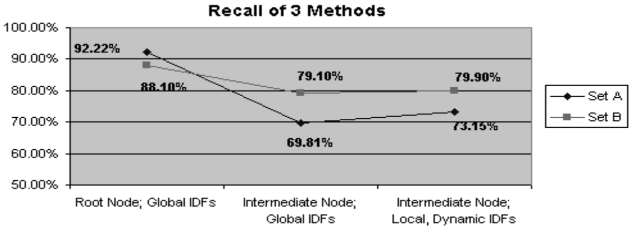

Figure 3. *Search Mode Results (Recall)*[1]

6. CONCLUSIONS

Professional users would apply a narrower concept of *on point* relevance to their queries than those used for our judgments, given their contextual understanding of their information need. Thus the recall obtained for the restricted search modes likely represents a lower bound on this performance. As a result of these alternative, user-centric search modes, energy previously spent on time-consuming source selection searches can be more productively spent on actual document retrieval and analysis. Moreover, an effective browse+search paradigm provides users the ability to exert more control over their searches.

7. 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 TREC-6*, pages 169–206, Nov. 1997.

[2] S. Park. Usability, user preferences, effectiveness, and user behaviors when searching individual and integrated full-text databases: Implications for digital libraries. *JASIS*, 51(5):456–468, March 2000.

[3] P. Thompson, H. Turtle, B. Yang, and J. Flood. TREC-3 ad hoc experiments using the WIN system. In *Proc. of TREC-3*, pages 211–217. NIST, Nov. 1995.

[4] W. Wang, W. Meng, and C. Yu. Concept hierarchy based text database categorization in a metasearch engine environment. In *Proc. of WISE '00*, June 2000.

[1] This activity resulted in two query subsets, one for tax and employment queries (Set A), the other for general queries (Set B).

[2] WIN stands for Westlaw is Natural.