Fall 2000

Better Search Engines for Law

Deanna Barmakian

Better Search Engines for Law*

Deanna Barmakian**

Ms. Barmakian compared fifteen search engines for effectiveness in retrieving legal information on the Web. She reports the results of two separate studies in which she examined known item searching and topical searching. She also considers the impact of various search engine issues on legal researchers.

¶1 With the increasing amount of legal information available on the Web, determining the efficacy of Web search engines in retrieving legal information becomes ever more critical to the legal researcher. This article describes two studies comparing the performance of search engines for legal research and then addresses various search engine issues and their impact on legal researchers. Several specific questions are considered:

- Given that there are specialty search engines for law, do these legal search engines provide more relevant results for legal research than general engines?
- If legal researchers are searching for a known item, which search engine should they use? Which should be used for topical searching?
- Given that only a small fraction of the Web is indexed, does this imply that only a similar small fraction of legal information on the Web is indexed? Or are legal and governmental information disproportionately represented in search engine indexes?

¶2 The first part of the article describes a study comparing the effectiveness of fifteen Web search engines in retrieving law-related known items. The second part describes a study comparing the effectiveness of ten search engines in retrieving items in response to legal topical queries. The third part examines various search engine issues—such as commercialization, spamdexing, the invisible Web, specialization, emerging technologies, and shared services—and their impact on legal researchers.

Known Item Retrieval

¶3 When legal researchers use the Web, they are, more often than not, looking for a known item: an agency publication, a government entity, a court, a searchable code, an organization, a publication by a research center, etc.

** Reference Librarian, Harvard Law School Library, Cambridge, Massachusetts.

¶4 Unfortunately, legal researchers often avoid *searching* for known items on the Web fearing they will be forced to sift through a long list of seemingly random search results. Rather than use a Web search engine to locate an item, legal researchers more typically guess at the URL, use bookmarks, or browse a promising gateway site. These hunt-and-peck strategies are cumbersome and inefficient. Finding a law-related known item would be more efficient if legal researchers knew which search engines were best at retrieving such items. The aim of the following study was to determine just that.

¶5 The study compared the efficacy of fifteen Web search engines in retrieving fifty law-related known items. These fifty items were searched for with an identical query on each engine within three hours of one another. Retrieval performance was ranked based on six categories of placement within, or lack of appearance within, the first twenty results.

Design of Known Item Study

Selection Process for Search Engines

¶6 Fifteen search engines were selected for inclusion in the study: AltaVista, Excite, Fast, Go's Infoseek, Google, HotBot, LookSmart, Lycos, Northern Light, Snap, WebCrawler, Yahoo!, GoTo.com, AltaVista's LawRunner, and FindLaw's LawCrawler.[1] The first thirteen are general interest engines and the last two, LawRunner and LawCrawler, are specialized engines for law.

¶7 The thirteen general search engines were selected[2] because they are heavily used search engines as indicated by meeting at least two of the following six criteria: ranked among top Web properties by Media Metrix,[3] ranked among top Web properties by Nielsen's NetRatings,[4] used in the Lawrence and Giles search engine studies,[5] mentioned on Search Engine Watch or Search

---

1. For a list of search engines used with their URLs, *see infra* appendix A.
2. The search engine selection process took place during September 1999.
3. Media Metrix monitors and records activity on the computers of several thousand people, a survey sample supposedly representing the entire Web community. *See Media Metrix, Media Metrix Top 50 US Web and Digital Media Properties for June 2000* (visited Aug. 1, 2000) <http://www.mediametrix.com/data/thetop.jsp>; Mick O'Leary, *Web Measurers Wrestle With Methodologies, Each Other,* ONLINE, May 1, 1999, at 105, 105.
4. Nielsen NetRatings, like Media Metrix, is also prominent in the field of Web measurement. *See* Neilsen NetRatings, *Top 10 Web Properties: Month of June 2000, U.S.* (visited Aug. 1, 2000) <http://63.140.238.20/hot_off_the_net.asp?country=north+america>; O'Leary, *supra* note 3, at 105.
5. Steve Lawrence & C. Lee Giles, *Accessibility of Information on the Web,* NATURE, July 8, 1999, at 107 [hereinafter Lawrence & Giles, *Accessibility of Information*]; Steve Lawrence & C. Lee Giles, *Searching the World Wide Web,* SCIENCE, Apr. 3, 1998, at 98 [hereinafter Lawrence & Giles, *Searching the World Wide Web*].

Engine Showdown,[6] used in previous relevancy ranking studies, or receiving coverage in the news.[7]

¶8 Legal search engines, engines specialized for law, are those whose indexes are created by crawling and indexing[8] sites known to contain legal material; LawCrawler (http://lawcrawler.findlaw.com), GOVBOT (http://ciir2.cs.umass.edu/Govbot/), and LawRunner (http://lawrunner.com)[9] are three that were considered for inclusion in the study.[10]

¶9 Legal directories are sites that contain categorized lists of handpicked legal Web sites. There are many of these, but to be considered for inclusion, legal directories had to have internal search engines that searched the titles and descriptions, if any, of the sites available in the directory. CataLaw (http://www.catalaw.com), RomingerLaw (http://www.romingerlegal.com), and the Internet Legal Resource Guide (http://www.ilrg.com) fit this category and were considered for inclusion.

¶10 Out of the legal engines and directories considered, only LawRunner and LawCrawler were included in the study. They performed the best on initial tests that consisted of ten searches for specific known items.[11]All of the legal directory sites performed too poorly for inclusion.[12] GOVBOT was initially included in the study but was later removed because of technical problems and poor performance.[13]

---

6. Search Engine Watch and Search Engine Showdown are two excellent Web sites that monitor and compare search engines. Danny Sullivan, *Search Engine Watch* (visited Feb. 20, 2000) <http://www.searchenginewatch.com>; Greg Notess, *Search Engine Showdown: The User's Guide to Web Searching* (visited Feb. 20, 2000) <http://www.notess.com/search>. *See* Bill Mickey, *A Web Search Trifecta: Keeping Tabs on Search Engine Features and Technology,* ONLINE, May 1, 1999, at 79.
7. Several searches were run in full-text news sources on Westlaw and LEXIS-NEXIS for articles about search engines published between September 1998 and September 1999. Headlines of news stories were scanned for mention of specific search engines.
8. A search engine index is a database of Web pages collected by pieces of software, often called robots or spiders, which "crawl" the Web from link to link, collect Web pages, and do various levels of indexing of the metatags and text of the page. In general, legal search engine creators release robots on government, military, legal academic, and commercial legal Web page servers rather than releasing robots that crawl the Web randomly.
9. LawRunner's index is not actually created by crawling and indexing legal sites. It runs on the AltaVista index but layers "behind the scene" Boolean searches on top of user queries to cull out legal material from the AltaVista index.
10. Unfortunately, two very good search engines specialized for government information, *Google Uncle Sam* (visited Jan. 23, 2000) <http://www.google.com/unclesam> and *SearchGov.com* (visited Feb. 19, 2000) <http://www.SearchGov.com>, were not yet available when the selection process was completed.
11. The test searches were for known items such as home pages of elected officials, agency pages, a treaty, etc. LawCrawler and LawRunner were selected because they performed markedly better than the other 4, locating the known item in 5 out of 10 searches.
12. Out of 10 searches, none of the legal directories located the known item in more than 2 searches, and they often produced no results at all.
13. After 12 searches in the study, GOVBOT had located only 1 of the items in its top 10 results. There were also repeated problems in connecting to the site.

¶11 Meta search engines were not considered for inclusion. Meta search engines allow users to send a search to several search engines at once.[14] Dogpile (http://www.dogpile.com), InferenceFind (http://www.infind.com), and Metacrawler (http://www.metacrawler.com) are examples of these. They perform no independent searching; they merely report the results of other engines. Most meta search engines garner results from engines already included in the study. For this reason, it seemed redundant to analyze them as well.

Selection Process for Known Items

¶12 The known items used in the study were selected at random from categories considered most useful for legal researchers: cases; searchable codes and rules; reports; e-serials; digitized collections; publishers; and the home pages of associations, councils, organizations, institutions, centers, courts, persons holding public office, federal and state government agencies, and professors.[15]

¶13 No known items were selected by searching with, or browsing the directory of, any Web search engine included in the study. Known items were selected using bookmarks, browsing forward from RomingerLegal (http://www.romingerlegal.com), WashLawWeb (http://washlaw.edu), the Legal Information Institute (http://www.law.cornell.edu), and from browsing legal research guides on the Web.

Query Design and Default Search Modes

¶14 This study sought to address what a typical legal researcher would retrieve using an informed, but not painstakingly designed, search. It did not try to reproduce the experience of an expert searcher using advanced features of search engines with carefully crafted searches. Consequently, the searches used in the study to find known items were fairly simple.[16] The level aimed for was something like "a legal researcher's first try," less than expert but above average. The only Boolean operator used was "and." Quotation marks were used in most cases for phrase searching. For instance, a search for a known item such as the home page of the National Bankruptcy Review Commission would simply be "national bankruptcy review commission."

¶15 Although a recent study determined that most Web users do not use multiword searches[17] and almost never use quotations or "and" in their searches,[18] the

14. For an excellent explanation of meta search engine strengths and weaknesses, along with descriptions of popular ones, see Nicholas Tomaiuolo, *Are Metasearches BETTER Searches?* SEARCHER, Jan. 1, 1999, at 30.
15. For a list of the known items used in the study, *see infra* appendix B.
16. For the searches used to find the known items, *see infra* appendix B.
17. In a study of 51,473 searches by 18,113 Excite users, 31% of the searches consisted of only one search term, and 67% of the searches consisted of two words or less. Bernard Jansen et al., *Real Life Information Retrieval: A Study of User Queries on the Web,* SIGIR FORUM, Spring 1998, at 5, 9.
18. In the same study "and" was used by only 5% of the searchers, and "+" or "-" or quotation marks were used by only 6% or less of the searches. *Id*. at 10.

searching behavior of the average Web user can hopefully be distinguished from that of the average legal researcher who should have more sophisticated online searching skills. For this reason, quotation marks and "and" were used in the searches.

¶16 The default search screen of each engine was used for all searches. In an effort to compare engines in their most often-used form, no advanced search screens or advanced features of the engines were exploited. This decision was based on the hypothesis that the average legal researcher usually does not take time to use advanced search screens or the features available on them. Moreover, the process of selecting which advanced features on these screens to use on any given search, on any given search engine, is a highly subjective process. Using default search screens would allow the study to address the more typical legal researcher, as well as reduce subjectivity.

¶17 The use of default search modes without exploiting any advanced features may have unfairly impacted the performance of many engines. For instance, LawRunner performs better if the pull-down menu listing jurisdictions or document types is used. Northern Light's performance improves when the file folders of results are explored. HotBot offers many advanced features, such as domain specifying, which, if used, enhance results. FindLaw's LawCrawler does not search the FindLaw directory, which, if searched, sometimes enhances performance. This is an unfortunate but necessary effect of using default search modes.

¶18 Following in this vein, the study used the same search on all search engines even though using customized search syntax might have improved performance for some engines. An experienced search engine user is aware of the specific syntax required by an engine for Boolean searching, phrase or proximity searching, URL and domain specifying, etc. Accordingly, such a user might be able to improve any of the search engines' performance in this study by using the specific "commands" required by a search engine to accomplish this. Nevertheless, using a standardized search for all search engines probably more accurately reflects the behavior of the average legal researcher.

Evaluation of Result Lists for Known Item Retrieval Performance

¶19 The top twenty items on search result lists were reviewed to determine whether or not the item had been retrieved. Rather than using strict placement among the top twenty to rank performance, a ranking system was devised to compare retrieval performance among the search engines. Known items had to appear within the first twenty items in result lists, or within one link of those twenty items. Result lists were analyzed and ranked using the following six categories:

- The item is among the first ten items (six points).

- An internal page of the item is among the first ten items (four points).[19]
- The item is within one link of an item in the first ten items (three points).[20]
- The item is among the eleventh through twentieth items (two points).
- The item is within one link of any of the eleventh through twentieth items (one point).
- The item is not found in any of the above categories (–two points).

The highest possible score was three hundred (six points times fifty searches).

Caveats

¶20 The study examined the first twenty results without regard to duplicate and inactive links. The aim was to punish engines that delivered duplicate hits and dead links.

¶21 Directory (also known as category or guide) hits counted in the top twenty results. For instance, if the first five items in the result list were from a directory, then those five directory links and fifteen results would be analyzed. This worked to the benefit of engines with strong directories and the detriment of those with poor directories. Links to items in Northern Light's Special Collections, its fee-based document delivery service, counted in the twenty results but were not counted as the correct item.

Results of Known Item Study

¶22 Table 1 lists the fifteen search engines ranked from best to worst performance in locating law-related known items on the Web.[21]

¶23 Table 2 presents more detail about the performance of these search engines in retrieving known items, noting the number of times each engine retrieved an item within one of the various ranking categories.[22] The total for all six categories for each search engine is fifty, the number of searches performed in the study. Multiplying the points per category times the number of instances the engine retrieved the item in that category yields the scores presented in table 1.

---

19. For this study, a Web page was considered an internal page if its URL could be truncated to locate the item. For instance, in searching for Sen. Diane Feinstein's official home page, a page titled *Diane Feinstein's Press Releases* that was a page *within* her official site, with a URL that was an extension of her official page's URL, would be considered an internal page.
20. It had to be a reasonable assumption that a page would contain a link to the correct item based on its title. For instance, in searching for Sen. Diane Feinstein's official home page, a result titled *U.S. Senators' Homepages* creates a reasonable assumption that it will link to Feinstein's page. The *U.S. Senators' Homepages* site's links would then be checked for an active link to Diane Feinstein's home page, the known item. In effect, this rewarded search engines that retrieved good list, or hub, pages.
21. It should be noted that after the study was completed, Lycos announced that it would begin delivering results from Fast rather than from its own search engine. *See Lycos.com Search Grows to Top Spot in Web Coverage—Surges in Search Popularity; Lycos.com Deploys Fast Search & Transfer Technology,* Bus. Wire, June 14, 2000, *available in* LEXIS, News Library, BWire File.
22. *See supra* ¶ 19.

Table 1
Search Engine Performance in Known Item

Rank	Search Engine	Known Item Retrieval Score (out of possible 300)
1	Yahoo!	229
2	HotBot	206
3	GoTo.com	200
4	Google	194
5	Snap	192
5	LookSmart	192
7	Fast	189
8	AltaVista	173
9	Northern Light	164
10	Excite	152
11	Go's Infoseek	145
12	Lycos	141
13	WebCrawler	109
14	LawCrawler	105
15	LawRunner	93

¶24 Two interesting aspects of these results: contrary to expectations, the legal engines performed much worse than the general search engines; and a surprising number of the search engines found the known item in the first ten results.

¶25 Based on the results of the study, legal researchers who need to locate a known item on the Web should have fairly good luck using a search engine, and they should probably use a search engine such as Yahoo!, HotBot, GoTo.com, Google, Snap, or LookSmart.

Implications for Percentage of Legal Information Indexed

¶26 The results of the study indicate that legal information on the Web is well indexed by search engines. Given a recent study reporting that only 16 percent of the Web was indexed,[23] this is surprising but welcome news for legal researchers. According to that result, out of fifty known items sought in the present study, only eight items should have been found. All fifty items, however, were found. Moreover, seven search engines put the known item in the top ten results more than 50 percent of the time.[24]

23. Lawrence & Giles, *Accessibility of Information, supra* note 5, at 107.
24. The known item was among the top 10 search results from Yahoo! 66% of the time; Google and LookSmart, 60% of the time; Snap, 58% of the time; HotBot, 54% of the time; and GoTo.com and Lycos, 52% of the time.

Table 2
Summary Table for Known Item Retrieval

	Ranking Categories*						Total
Engine	**Cat. 1**	**Cat. 2**	**Cat. 3**	**Cat. 4**	**Cat. 5**	**Cat. 6**	**Score**
Yahoo!	33	0	13	0	0	4	229
HotBot	27	2	15	0	1	5	206
GoTo	26	2	16	0	0	6	200
Google	30	1	9	0	1	9	194
Snap	29	2	9	0	1	9	192
LookSmart	30	1	8	0	2	9	192
Fast	20	7	14	2	3	4	189
AltaVista	21	4	15	0	2	8	173
Northern Light	16	9	14	0	4	7	164
Excite	18	5	13	1	3	10	152
Go's Infoseek	24	3	7	0	0	16	145
Lycos	26	3	3	0	0	18	141
WebCrawler	15	4	13	0	0	18	109
LawCrawler	7	10	16	0	3	14	105
LawRunner	7	5	19	0	4	15	93

* Category 1: in first 10 items (6 pts.)
Category 2: internal page first 10 items (4 pts.)
Category 3: w/in one link of first 10 items (3 pts.)
Category 4: in 11th–20th items (2 pts.)
Category 5: w/in one link of 11th–20th items (1 pts.)
Category 6: not found (-2 pts.)

¶27 The legal search engines were the poorest at retrieving known items in their top ten results. Even so, they both retrieved the items in their top ten at least 14 percent of the time.

¶28 Accordingly, legal researchers should not shy away from using search engines to locate law-related items on the Web. For instance, in trying to locate a GAO report, the study indicates that researchers should simply search for the report's title on Yahoo! rather than find the GAO home page and use a search facility there. The items legal researchers seek seem to be exactly those most likely to be indexed by search engines.

Performance Observations

¶29 Observations were made about engine retrieval performance based on types of documents searched for and types of searches used. Conclusions cannot be drawn, however, since different styles of searching were used to locate similar types of documents. For example, multiphrase searches stump the retrieval algorithms of certain engines. Those same engines may excel at retrieving agency documents

with simple searches. However, if multiphrase searches were generally used to locate agency documents in the course of this study, those engines would be assessed unfairly in their performance in retrieving agency documents. Nonetheless, here are some comparative observations that may be of interest to legal researchers.

¶30 Comparing performance by types of searches used:

- AltaVista, GoTo.com, LookSmart, and Yahoo! were best for two-phrase searches,[25] such as "national conference of commissioners" and "uniform state laws";
- Fast and Google were best for complex searches[26] using phrases and several words such as "john h jackson" and law professor and michigan;
- Snap, Yahoo!, and Lycos were best for exact phrase searching,[27] such as "nebraska state bar association";
- Lycos was best for multiword searches without connectors or phrases,[28] such as los angeles municipal code.

¶31 Comparing performance by types of documents:

- Google, LookSmart, Snap, and Yahoo! were best for finding associations, councils, commissions, organizations, agencies, and departments;[29]
- Google and Infoseek were best for finding the top page of serial publications;[30]
- Google, Lycos, and Snap were best for finding the top page of rules or codes online;[31]
- Fast and Google were best for finding a full-text document, nonprimary law, i.e., reports, articles;[32]
- Fast was best for finding a full-text document, primary law, i.e., federal and state cases, agency regulation, public law, etc.[33]

Directory/Category Results

¶32 It is common for search engines to deliver directory, or category, entries ahead of Web pages in search result lists. Depending on the quality of the directory, this can aid or hinder a researcher looking for a known item.

---

25. Out of 8 such searches, these 4 search engines found 5 of the items in their top 10 results.
26. Out of 7 such searches, Fast found 6 and Google found 5 of the items in their top 10 results.
27. Out of 25 such searches, Snap found 20, Yahoo! found 19, and Lycos found 18 in their top 10 results. Most other engines found 15 or less out of 25.
28. Unfortunately, this was a category of only 3 searches. Lycos, however, was the only search engine to find all 3 items in its top 10 results.
29. Out of 13 such searches, Yahoo! and Snap found 11 of the sought items in their top 10 results. Google and LookSmart found 10 of the sought items in their top 10 results.
30. Out of 5 such searches, Google and InfoSeek found the sought item 4 times in their top 10 results.
31. Out of 6 such searches, Lycos, Snap and Google found the sought item 5 times in their top 10 results.
32. Out of 6 such searches, Fast and Google found the sought items 5 times in the top 10 results.
33. Out of 9 such searches, Fast found the sought item 7 times in its top 10 results. Other engines found the item 4 or less times in their top 10.

¶33 For instance, Yahoo! and LookSmart both have well-developed, human-editor-created directories of Web sites. In using Yahoo! and LookSmart to search for known items in this study, the items often appeared in their "directory" hits. However, several engines were hurt by the number of directory links included in search result lists. This was true for Snap, Excite, and, especially, HotBot. The "hits" from their directories rarely, if ever, led to the queried item.

Need for Further Research on Known Item Retrieval

¶34 The methodology used to perform this study was simplistic. Searches were done by hand without the use of automated software or agents. Only fifty items were searched for out of the many millions of items on the Web. Searches were designed for the study and not culled from a pool of "real" searches in a legal setting. Consequently, assessment and observations must be considered anecdotal rather than conclusive.

¶35 Rather than a definitive work, this study should be a starting point for further research. More work needs to be done in assessing how well or poorly Web search engines retrieve items known to exist on the Web. Legal researchers and librarians do this type of Web searching often. More research on known item searching would perhaps alert search engine developers to the issues involved in this prevalent form of searching.

Conclusions for Known Item Searching

¶36 The study on using search engines for retrieval of law-related known items on the Web indicates that

- Despite the low percentage of the Web that is indexed, search engines are effective in locating law-related known items;
- Yahoo!, HotBot, GoTo.com, Fast, Google, Snap, and LookSmart are the better search engines to use for law-related known item searching; and,
- Despite being specialized for law, legal search engines should not be used for law-related known item retrieval.

Armed with this information, legal researchers should be able to find legal materials on the Web more efficiently by using search engines instead of hunt-and-peck techniques, and by using the better search engines for this type of searching.

Topical Information Queries

¶37 Traditionally, legal researchers have relied on commercial resources to find background information on a legal topic. Increasingly, however, they are turning to the Web as a "no-cost" (or lower cost) alternative. For certain types of

information, such as foreign law, and foundation, association, and company information, the Web is viewed by many as a better approach than the traditional legal research sources.

¶38 The major drawback for using the Web for topical information, however, is the inefficiency of search engines in retrieving relevant material. Search engine indexes are vast. Simple topical queries retrieve thousands of results. Even with sophisticated ranking algorithms, search engines do not always deliver the most relevant results at the top of the retrieval lists. Although this is true for all search engines, it may be that some engines deliver more relevant results than others. If legal researchers knew which engines delivered the most relevant results for law-related subjects, they could make an informed decision when selecting a search engine for topical searches.

¶39 To try to answer this question, the following small-scale study was undertaken to compare search engine performance in delivering relevant results for law-related topical searches.[34]

Design of Relevancy Study

Selection Process for Search Engines

¶40 Ten engines were selected for inclusion in the study: FindLaw's LawCrawler, AltaVista's LawRunner, Yahoo!, HotBot, GoTo.com, Fast, Google, Snap, LookSmart, and AltaVista. These were a subset of the fifteen engines used in the known item retrieval study described earlier in this article.[35] The number ten was arbitrary. The intention was for the study to be reasonable in scope and to pit some of the top general search engines against legal search engines.

Selection Process for Topical Information Searched

¶41 Ten searches were used in comparing the relevance of results.[36] Five searches were based on questions asked at the Harvard Law School Library reference desk.[37] The rest were devised to "round out" the types of information sought in

---

34. Topical searches are distinct from known item searches in that the researcher does not know if Web pages exist on the topic. For instance, a researcher interested in gun owners' rights might seek a known item such as the National Rifle Association home page. Perhaps fearing that the information on such a site would be biased, the researcher may want to run a topical search on gun control or gun owners' rights generally to see what kinds of information exist on the Web.
35. The original 15 engines were selected for inclusion in the study if they were at least two of the following in September 1999: mentioned on Search Engine Watch or Search Engine Showdown, used in previous relevancy-ranking studies, used in the Lawrence & Giles studies, ranked among top Web properties by Media Metrix or Nielsen's NetRatings, or receiving coverage in the news. *See supra* ¶ 7.
36. For a list of the relevancy searches and the types of information considered to be relevant for each, *see infra* appendix C.
37. During September and early October 1999, two librarians kept a list of topical reference questions for which they used a search engine in the process of helping a patron. Five topics from this list were selected for use in the study.

the study. In devising the searches, consideration was given to the types of information legal researchers look for on the Web rather than solely in commercial online or print legal sources.

¶42 All searches were law-related. One search was for regional information (lawyer referral services in Chicago.) Three searches sought information about legal entities (a legal philosopher, a defunct federal commission, and an existing state commission.) Two searches involved international or foreign legal information (a convention on North Atlantic salmon and freedom of religion in Guatemala). Two searches sought broad topical information (discovery rules for electronic evidence, and licensing and certification programs for midwives). One search was for information about the adjudicative process in the Social Security Administration. Finally, one search was for information about a law review, the *UCLA Bulletin of Law and Technology.*

Query Design and Default Search Modes

¶43 As with the searches designed to retrieve known items, the topical searches were not painstakingly designed.[38] The search level aimed for was "a quick try by a legal researcher." Searches were all at least two words, and the only Boolean operator used was "and." Quotations were often used for phrase searching.

¶44 The same query was used on each engine in its default mode. Advanced features and customized search syntax were not used. For each topical search, the ten engines were searched within a four-hour time span on the same day.

¶45 As in the known item study, the aim was *not* to reproduce the experience of someone intimately familiar with the particular syntax and advanced features of search engines.[39] Instead, the study compared engine performance from the viewpoint of an average legal researcher who uses search engines but doesn't make a habit of exploiting the unique features of each. As a result, the study does not answer the question of how the engines would compare if each were searched by an expert using the advanced searching capabilities of each engine.

Scoring the Relevancy of Result Lists

¶46 The top ten results from each engine were evaluated and assigned zero, one, two, or three points. Zero points were awarded to dead links, false hits, duplicate hits (with the exact URL), and pages on which the search terms were only mentioned. One, two, or three points were awarded based on degree of relevancy and amount of information provided.[40]

¶47 For each search, a scoring system was devised based on the information sought. The scoring system defined the length of the discussion and the type of

38. *See supra* ¶¶ 14–15.
39. *See supra* ¶¶ 16–18.
40. *See infra* appendix C for a list of searches and scores awarded to types of information retrieved.

information that would receive a certain score. Scoring on a predefined model reduced the potential for subconscious favoring of one engine's results over another's.

¶48 As an example of how results were scored, consider the search for information on the National Commission on Reform of Federal Criminal Laws. Three points were awarded to pages containing more than a paragraph of information about the commission or the full text of its final report. Two points were awarded to pages containing at least a paragraph about it or a link to the text of its final report. One point was given to pages containing citations of its publications or a sentence about it.

¶49 The pages retrieved were closely tracked. Once a page from a result list was evaluated and scored, it received the same score when retrieved by any other engine. This reduced any subconscious predilection to favor the results of one engine over another. Arguably, any relevance evaluation is subjective. The same person, however, evaluated each of the 1,000 pages in the result lists. This at least reduced, if not eliminated, the problem of disparate notions of relevancy.

¶50 Pages in result lists were evaluated individually. The overall content delivered in the list was not scored. That is, if an engine returned eight unique pages containing the same information worth one point, plus two pages worth zero points each, it received a score of eight. It wasn't penalized for redundancy of information returned.

¶51 The highest possible total score for the study was three hundred: three points times ten results times ten searches.

Results of Relevancy Study

¶52 Table 3 lists the ten search engines in order of best to worst performance in retrieving pages relevant to legal topical queries.[41] Interestingly, the search engine ranking for topical searching does not correlate with search engine performance for known item retrieval.[42] Surprisingly, the legal search engines outperformed all the general search engines except Google. In contrast, for known items, they were outperformed by all of the general search engines.

¶53 Yahoo! dropped from top performer in known item retrieval to one of the worst performers for relevance of topical search results. One cause was that on several searches Yahoo! returned two or three highly relevant directory entries, but

41. On June 26, 2000, after the study was completed, Yahoo! announced that it would use Google as its default "behind the scenes" search engine for its Web page results rather than Inktomi. *See* Alan T. Saracevic, *Yahoo! Clicks on Google as Primary Search Engine; Deal a Blow to Inktomi, Whose Shares Tumble,* S. F. EXAMINER, June 26, 2000, at C4. Had Yahoo! been using Google during the study, its scores for relevance of results for legal topic queries would certainly have been affected for the better. However, it is important to note that the "category" or "directory" items at the top of Yahoo! result lists would have been the same, somewhat ameliorating the effect on the study results of this drastic change.

42. *See supra* ¶ 22.

Table 3
Search Engine Performance in Relevant Retrieval

Order of Performance	Search Engine	Relevance Score (out of possible 300)
1	Google	149
2	LawCrawler	120
3	LawRunner	111
3	Fast	111
5	LookSmart	101
6	AltaVista	98
7	Snap	95
8	HotBot	92
9	Yahoo!	90
10	GoTo.com	87

no other Web pages. On these searches it received low scores even though it had delivered relevant items. It was perhaps unfairly punished for not delivering at least ten results.[43] Otherwise, the shift from the top to the bottom is a perplexing surprise.

¶54 Notably, scores for the relevance of search engine results for legal topical queries was low. Even the top performing search engines did not receive high relevancy scores. Out of a possible three hundred points, none of the search engines in the study received even half that amount. This does not bode well for legal researchers attempting to use the Web as an alternative to commercial resources when researching legal topics.

Overlapping Result Sets

¶55 Patterns emerged upon analyzing the results of the study. There was significant clustering in result sets from two groups of engines. One group, composed of Yahoo!, GoTo.com, Snap, and HotBot, often returned similar results. This is most likely due to the fact that these four engines were powered by Inktomi. They searched and delivered from the same Inktomi database, although each had some customization of the database and search features.

¶56 A second group, consisting of FindLaw's LawCrawler, AltaVista's LawRunner, AltaVista, and LookSmart, also produced similar result sets for several searches. This is not so easily explained since there were no shared directories

43. The design of the study was such that a search engine that returned 10 low relevance items (one point each) scored higher than a search engine that returned only 2 or 3 highly relevant pages (three points each).

or "behind the scene" services among these search engines except for the shared base index of AltaVista and LawRunner.

¶57 The following is a list of instances in which there was overlap of search engine results for law-related topical queries.

- Yahoo!, GoTo.com, Snap, and HotBot returned at least six of the same pages in three searches;
- GoTo.com, Snap, and HotBot returned at least four of the same pages in six searches;
- LawCrawler, LawRunner, and AltaVista returned at least five of the same pages in three searches;
- LawCrawler and LookSmart returned at least eight of the same pages in three searches; and
- LawCrawler and LawRunner returned at least seven of the same pages in five searches, at least five of the same pages in seven searches, and the same ten results for two searches.

¶58 The overlap between the Yahoo! group and the LawCrawler group was low.

¶59 Google, the best performer, and Fast, tied for third best, generally returned result lists unique from those returned by either of the two groups.

¶60 Consequently, legal researchers willing to use more than one search engine for topical research might want to use Google first, one from the LawCrawler group second, (LawCrawler, LawRunner, AltaVista, or LookSmart), Fast third, and one from the Yahoo! group fourth (Yahoo!, GoTo.com, Snap, or HotBot). A legal researcher who plans to use only one search engine for topical legal research may want to use either Google or FindLaw's LawCrawler.

Unique Result Sets

¶61 A result set was considered unique if it contained at least seven results not returned by any of the other nine engines for that topical query. However, uniqueness of results was analyzed completely separately from relevance of results; the relevance of a "unique" set could be low or high. Uniqueness is noted here because it may be of interest to those attempting to do comprehensive Web mining for topically related pages.

¶62 In six out of the ten searches used in this study, Google and Fast had unique result sets. That is, they both located at least seven pages not found by any of the other nine engines in six out of ten searches.

¶63 LookSmart also provided unique results, but not as often as Fast or Google. LookSmart delivered unique results in four out of ten searches.

¶64 Based on the results of this small-scale study, it may be worth searching both Google and Fast in addition to other engines. This would help researchers locate pages not likely to be found by other engines.

Conclusions for Relevancy Study

¶65 The study comparing the relevance of results delivered by search engines for legal topical queries indicates that

- legal search engines such as LawRunner and LawCrawler deliver more relevant results than general search engines for legal topic queries;
- there is significant overlap of results between the two legal search engines so time-pressed legal researchers should use just one or the other;
- there is overlap in results from engines that use Inktomi, so legal researchers should use just one from this group (Snap, HotBot, and GoTo.com);
- Google and Fast deliver "unique" results with little overlap of results from other search engines; and
- relevance of search engine results, even from legal search engines, is low, indicating that search engines are not yet viable alternatives to commercial legal resources for topical legal research.

Search Engine Issues and How They Impact Legal Researchers

¶66 Issues surrounding search engines are provocative, complex, interesting, and sometimes depressing. Even though search engines are becoming more sophisticated and powerful, their target, the Web, is less and less conquerable. Fortunately, some of the negative aspects of search engines do not affect legal researchers as much as they do general Web users. Moreover, some of the new developments actually aid searchers of law-related information. The following section surveys search engine issues, both positive and negative, and hypothesizes as to their importance, or lack thereof, to legal researchers.

Search Engines with Unique Ranking Mechanisms

¶67 Search engine relevancy-ranking algorithms are carefully guarded trade secrets and developers are loathe to part with specifics about them. At their most basic level, ranking algorithms take into account position of searched terms in a Web document, frequency of searched terms, and percentage of searched terms compared to total words in the document. Thankfully for legal researchers, a series of innovations in relevancy ranking have come to the fore, which, layered on top of typical ranking mechanisms, improve search engine performance. These new ranking mechanisms capitalize on the link topology of the Web, user popularity, and the registering of keywords.

Relevance Based on Link Topology: Google and the Clever Project

¶68 The Google page-ranking mechanism is innovative and effective. It ranks according to how many other pages link to a particular page, and how important those linking pages are. Each page in Google's index is assigned a "PageRank";

the more links to a page, the higher its PageRank.[44] Links from authoritative[45] sites, like Yahoo! for example, give an extra boost to a PageRank. The underlying, sound assumption is that a link to another page is an implicit endorsement of that page by a Web author. Google capitalizes on this editorial judgment of Web authors by delivering those pages most linked to, or "cited," by Web authors at the top of result sets.[46] In essence, Google uses a result-ranking system based on Web author endorsement of pages.

¶69 In addition to Web author endorsement, Google also takes anchor text references into account when ranking pages. Anchor text is the word (or words) that has been used as a link to a URL. Google developers assume that if several pages point to the same page using the same anchor text, the page they point to is probably relevant for those anchor text terms.[47] For instance, if there are hundreds of links to "IRS Tax Forms" using the same URL as the underlying link, a Google user who types in "IRS Tax Forms" will see that URL at the top of the result list.

¶70 These types of ranking mechanisms are highly advantageous to legal researchers who are often searching for those pages frequently cited (linked to) on the Web. They are also searching for sites consistently and popularly named, making anchor text references highly relevant for legal sites. Governmental pages, statutory codes, court pages, etc., are linked to heavily using the same anchor text (link text). In the studies described earlier in this article, Google performed extremely well. In fact, Google outperformed most other engines in searches for the Web pages of properly named entities such as agencies, associations, commissions, departments, rules, codes, and reports. It excelled at finding these law-related pages because of its use of Web author endorsements (link topology) and anchor text references in ranking results.

¶71 Not yet available as a public search engine, the Clever Project at IBM's Almaden Research Center is creating a search tool that also capitalizes on the link topology of the Web.[48] In response to a query, Clever generates a two-part list

---

44. The algorithm determining PageRank can be thought of as analogous to the behavior of a hypothetical Web surfer who starts at a random page and keeps following links, never hitting back. When bored, the surfer starts at another random page. The probability that the surfer will visit a page is its PageRank. *See* Sergey Brin & Lawrence Page, *The Anatomy of a Large-Scale Hypertextual Web Search Engine* (visited Aug. 22, 2000) <http://www7.scu.edu.au/programme/fullpapers/1921/com1921.htm>.
45. Google's creators, Stanford computer science graduate students Sergey Brin and Larry Page, spent three years analyzing more than a billion Web links to determine which were the authoritative sites on the Web. *See id*.
46. Since Google's release, many other search engines, such as AltaVista, Excite, Fast, Go, Inktomi, and Northern Light, have begun to use "link analysis" in their ranking algorithms. *See* Danny Sullivan, *Search Engine Features for Webmasters* (last modified June 1, 2000) <http://searchenginewatch.internet.com/Webmasters/features.html>.
47. *See* Greg R. Notess, *Rising Relevance in Search Engines,* ONLINE, May 1, 1999, at 84, 85.
48. As of April 1999, IBM was seeking to license Clever technology to portal sites and to organizations with large intranets. *See* David Green, *Search Insider,* INFO. WORLD REV., Apr. 1, 1999, at 19, 20.

containing "authorities" and "hubs." Authorities are the best Web pages on a topic, and hubs are Web pages with the best collections of links for a topic.[49] A good hub points to many authorities and a good authority is pointed to by many good hubs.

¶72 The Clever algorithm works something like this. For any query, Clever retrieves the first two hundred pages that AltaVista delivers for that query. Clever then augments that set by adding all pages that link to *and from* those two hundred pages, forming a collective group of pages (generally between one thousand to five thousand pages) called the root set. For each page in the root set, Clever first assigns random authority and hub scores. (The authority score is the sum of the scores of the hubs that point to it. The hub score is the sum of the authority scores to which it points.) It reassigns scores in an iterative process, about five rounds for a root set of three thousand, until it generates final authority and hub values and delivers the highest scoring authorities and hubs in results.[50] Researchers on the project believe that Clever-created lists of Web resources for a topic compete favorably with the human-compiled guides of Yahoo! and Infoseek.[51]

¶73 Legal researchers should keep an eye out for announcements of licensing agreements by major search engines to use Clever technology. Clever, like Google, seems well suited to the needs of academic and legal researchers. Although slower than Google since it does not have its own index and must run through gathering and ranking steps before delivering results, it may be more effective than Google. Clever looks both backward *and forward* from starting pages and creates lists of "gateways" or "hubs" on a topic, a task traditionally the domain of librarians and information professionals. If Clever performs as promised by its creators, it may obviate the need for law librarians to create these lists by hand for patrons or for library Web pages.

Relevance Based on User Behavior: Direct Hit

¶74 Direct Hit[52] is a both a search engine and a licensed technology that ranks pages according to their popularity among users of search engines. Direct Hit tracks the sites previous searchers have selected from search engine result lists and notes how long they stay at the selected site. The more the site is selected from a result set for a certain search, and the longer the searcher stays, the higher the page will be ranked for the next user with a similar search.[53]

---

49. *See* David Gibson et al., *Structural Analysis of the World Wide Web* (visited Aug. 11, 2000) <http://www.w3.org/1998/11/05/WCworkshop/Papers/kleinber1.html>.
50. *See* Soumen Chakrabati et. al., *Hypersearching the Web,* SCI. AM., June 1999, at 54, 58.
51. *See id*. at 60.
52. *Direct Hit* (visited July 26, 2000) <http://www.directhit.com/>.
53. *See* Rex Crumb, *Direct Hit Gets Searches on Target With HotBot,* 19 B. BUS. J., Mar. 19, 1999, at 6; Bobbi Cross & Michelle Ayers, *Internet Searching Advancements,* LEGAL INTELLLIGENCER, Dec. 2, 1998, at 9.

¶75 Direct Hit technology is licensed to HotBot, Lycos, Infoseek, and LookSmart, among others.[54] Some, such as Lycos, HotBot, and Infoseek, simply deliver Direct Hit results as the first ten items in their result lists. Others, such as LookSmart, do not merge Direct Hit results into regular search results but instead provide a link to Direct Hit results (e.g., "Top Ten Most Visited Sites for _____") at the top of result lists.

¶76 Unfortunately, popularity-based ranking mechanisms have the potential to become a self-fulfilling prophecy over time. (Popularity is determined by how often people select a site from a result list. Result lists only deliver popular sites. People select from result lists. What they select becomes popular.) Moreover, Direct Hit results probably favor older sites over newer ones.

¶77 Regardless of this drawback, testing on Direct Hit revealed that basic law-related searches, such as "montana law" or "barbara boxer," work very well. Results are highly relevant. More in-depth legal searches do not always deliver such relevant results, but, in general, Direct Hit results for legal searches are quite good. HotBot's strong performance in the known item study described earlier is evidence for this, since HotBot result lists begin with the top ten results from Direct Hit. Although HotBot was one of the top performers for known item retrieval, it did not do well in retrieving information relevant to a legal topic, ranking eighth out of ten. One possible explanation is that the legal topical queries used in the study were unique, containing unusual combinations of search terms. Direct Hit may have had a harder time ranking results for such searches because it lacked enough data from previous searchers. Consequently, legal researchers using uncommon search terms may want to avoid popularity-based search engines such as Direct Hit.

Retrieval by Registered Keywords: RealNames

¶78 RealNames,[55] owned by Centraal Corp., is not actually a search engine. It is a technology licensed by many search engines, ISPs, and major browsers. RealNames' goal is to displace URLs by allowing users to enter words ("Internet Keywords") instead of URLs to go to a site. Thus, rather than type a URL in the address bar of Netscape[56] or Internet Explorer, a user can enter the name of a product, university, or company, etc. If the Web developer of the official site of

---

54. *See Direct Hit Partners* (visited Aug. 12, 2000) <http://directhit.com/about/partners/>. Direct Hit should not be confused with GlobalBrain technology, a popularity-based ranking mechanism licensed in January 2000 by Snap.com. Snap maintains a "LiveDirectory" to which site owners can add their site. As "Live Directory" sites develop traffic, their rankings in Snap search results rise, eventually pushing them to the main Snap directory. *See* Duane Allen, *Product Has Link to South,* SOUTHLAND TIMES (New Zealand), Jan. 4, 2000, at 9.
55. RealNames, *Internet Keywords* (visited July 26, 2000) <http://Web.realnames.com/>.
56. For Internet Keywords to substitute for URLs on Netscape, users must first download a free piece of software called IKTune-Up from RealNames. *See* RealNames, *What is IKTune-Up?* (visited July 26, 2000) <http://Web.realnames.com/Virtual.asp?page=IKTuneUp_Download>.

that product, university, or company has registered with RealNames, the user will be taken immediately to that site. For instance, on Netscape or Internet Explorer, a user can enter "brown university" or "ford" in the address line and be taken instantly to the official home page of Brown or Ford without having had to guess the URL or run a search.[57]

¶79 Search engines utilizing this service include AltaVista, LookSmart, Go's Infoseek, and Google. Users who enter "Internet Keywords" as search terms in any of these search engines will see a link at the top of search results to a RealNames result, usually with a superscripted RN symbol next to it.

¶80 Legal researchers are not necessarily heavily impacted by this development but do benefit from it. This technology makes it easier to locate the official pages of companies, federal agencies, and some educational institutions.[58] Not all Web developers of legal sites have registered their keywords with RealNames, however. Legal entities such as state agencies, bar associations, and law libraries are unlikely to have yet registered their keywords. Over time, more of them will.[59] Consequently, RealNames is an interesting development but of no immediate importance for legal researchers. Of far more importance is the advent of specialty search engines for law.

Specialty Engines

¶81 Specialty engines create their indexes by crawling only a hand-selected set of sites, sites on a certain domain (i.e., .gov, or .mil), or by only indexing pages containing certain types of files. Creators of specialty engines aim to improve relevancy by creating an index that is focused on a particular topic, region, language, or information type, such as news, MP3 files, or shopping information.[60]

---

57. Registrants with RealNames cannot unfairly commandeer keywords. RealNames verifies that registrants have a trademark for the keywords they register. *See* Lee Pender, *Centraal Gets Personal with My RealNames,* PC WEEK, Apr., 19, 1999, at 32, 32; *Business is the Game of the Name,* ECONOMIST, July 31, 1999, at 59, 59.

58. In February 2000, for example, entering "federal judiciary," "environmental protection agency," and "library of congress" in the URL bar in Internet Explorer, users would be taken directly to the official home pages for these entities.

59. In February 2000, "massachusetts board of higher education" and "alameda county bar association" were not registered with RealNames. However, the names of some large law firms were, such as Hale & Dorr and Skadden, Arps, Slate, Meagher, & Flom. By July 2000, however, the Alameda County Bar Association (a randomly selected sample organization in February) had registered its name with RealNames. As of 1999, RealNames had already registered about 100,000 keywords. *See* Linda Leung, *Centraal, Now RealNames, Secures $70M Worth of Funding,* Newswire, Aug. 10, 1999, *available in* WESTLAW, NWSWIREVNU database, 1999 WL 6823407.

60. An example of a regional search engine, *Somewherenear* (visited Aug. 12, 2000) <http://somewherenear.com>, is a U.K. search engine allowing users to retrieve Web pages of a particular business type located near them by entering the business type and specifying their location. An example of a topical search engine, *Medical World Search* (visited Aug. 12, 2000) <http://www.mwsearch.com>, is specialized for medical information. Domain-specific search engines include *SearchMil.com* (visited Aug. 12, 2000) <http://www.searchmil.com>, searching only sites on the .mil domain, and *Canada.com* (visited Aug. 12, 2000) <http://www.canada.com>, searching only sites on the Canadian (.ca) domain. An example of a foreign language search engine, *Terra* (visited Aug. 12, 2000) <http://www.terra.es>, offers Spanish language results only. *OnlineNewspapers.com,*

Increasing numbers of specialty engines are being developed. Unfortunately, which, if any, outperform general engines is as yet largely untested.[61]

¶82 There are many search engines specialized for legal and government information. They include LawCrawler (http://lawcrawler.findlaw.com), LawRunner (http://lawrunner.com), GOVBOT (http://ciir2.cs.umass.edu/Govbot), Google Uncle Sam (http://www.google.com/unclesam), SearchGov.com (http://www.searchgov.com/), and USGovSearch (http://usgovsearch.northernlight.com), among others. As indicated by the study described earlier, using legal search engines improves the relevance of results for legal topical queries.

¶83 A true boon to legal researchers, however, would be the development of a combination regional-legal search engine. Imagine, for example, a search engine specific to a region, perhaps California or New York, which only searched legal and government sites. This would bring us a great deal closer to jurisdictional searching on the Web. As yet there is nothing like this,[62] but with the increase in specialty search services, perhaps one will be developed.

Shared Services and Directories

¶84 Although search engines often share underlying technology or directories, they do not produce the same result sets for the same search. The following is a brief discussion of this phenomenon.

Example of a Shared Service: Inktomi

¶85 Inktomi is a corporation that houses and powers the indexes of many major search engines, such as HotBot, GoTo.com, LookSmart, and Snap.[63] It formerly powered Yahoo! as well.[64] (Snap delivers Inktomi results when its directory fails to generate hits, as did Yahoo!.[65] HotBot and GoTo.com deliver Inktomi results as a matter of course.[66]) Search engines contract with Inktomi because it offers an immense database with sophisticated search features. Inktomi's forte is effective

---

<http://www.onlinenewspapers.com>, delivering only newspaper articles as results, is an example of a search engine specialized by information type.

61. There was, however, one interesting study comparing five search engines' capacity to retrieve information relevant to European Union research. One of the search engines was specialized for the topic. Ten searches were used and the top 20 results were evaluated based on their appearance of relevancy from the result list display. *Euroferret,* the search engine specialized for European Union information, performed worst. *See On Test: Web Search Engines,* EUR. INFO.: J. EUR. INFO. ASS'N, Apr. 1999, at 35, 36.
62. By using advanced searches (i.e., Boolean searches using "law or legal" and state name), and limiting the domain to .gov or .edu, it is possible on a limited scale to create the effect of jurisdictional searching on the Web.
63. *See* Danny Sullivan, *Search Engine Alliances Chart* (visited July 26, 2000) <http://www.searchenginewatch.com/reports/alliances.html>.
64. *See* Mylene Mangalindan, *Yahoo! Says It Is Switching to Google From Inktomi for Search Technology,* WALL ST. J., June 27, 2000, at B4.
65. *See* Greg Notess, *The Many Faces of Inktomi,* DATABASE, Apr. 1999, at 63, 64.
66. GoTo.com delivers paid advertiser results first, then Inktomi results. *See id.*

realization of parallel computing technology, which allows many cheap computers to work together to house millions of documents and search them in a quarter of a second.[67]

¶86 Although these search engines are all powered by Inktomi (their indexes sit on Inktomi hardware and searches are run with Inktomi software), they do not produce the same result sets for any given search.[68] This is because each Inktomi-powered engine has its own somewhat customized version of the Inktomi database;[69] search engines can have pages unique to their search service in addition to Inktomi pages. Inktomi customers can also customize which Inktomi advanced searching features they will offer their users. Inktomi offers full Boolean and phrase searching, truncation, language limits, and searching within title, specified domains, and specified hosts.[70]

Shared Directories

¶87 Directories, human-compiled categorized lists of Web pages, are sometimes shared between search engines. Yahoo!'s directory is the oldest, and most likely the largest, of the human- compiled directories of Web pages.[71] The Yahoo! directory is not licensed to other search engines. There are two human-compiled directories on the Web that compete with Yahoo!, the LookSmart directory and the Open Directory.

¶88 LookSmart has approximately 150 editors compiling and categorizing Web sites into its directory.[72] LookSmart licenses its directory to AltaVista, HotBot, and Excite as well as other major Web sites.[73]

¶89 AltaVista, besides using the directory from LookSmart, also uses sites from the Open Directory, a volunteer-created, immense directory of the Web.[74] Now owned by Netscape, the Open Directory is given away to several search

67. *See* Charles Piller, *Parallel Universe; Inktomi Has Built its Reputation as a Powerhouse Behind the Web Search Firms with its Redundant, Neutral System,* L. A. TIMES, Apr. 19, 1999, at C1.
68. In the relevancy study described earlier, 10 topical queries were used to compare the relevancy of results from 10 search engines. There was some overlap in results from the Inktomi-powered group of search engines. For every search, however, they each offered some unique pages in their top 10 results. *See supra* ¶ 55.
69. *See* Notess, *supra* note 65, at 63.
70. *See id*. at 64.
71. Yahoo! does not provide information about the size of its directory, but competitors estimated its size to be between 2 and 4 million sites in fall 1999. *See* Leslie Walker, *Wheat.com vs. Chaff.com,* WASH. POST, Oct. 28, 1999, at E01.
72. *See LookSmart Partners with IDG.net, Go2Net and Blue Mountain Arts; LookSmart Continues to Gain Momentum as a Preferred Search Content Provider,* Bus. Wire, Apr. 8, 1999, *available in* LEXIS, News Library, BWire File; Pete Barlas, *LookSmart's Big Ambition,* INVESTOR'S BUS. DAILY, Apr. 5, 1999, at A6.
73. *See LookSmart Gains Momentum as Category Search Leader; Three of Top 10 Web Sites Utilize LookSmart's Directory,* Bus. Wire, Feb. 9, 1999, *available in* LEXIS, News Library, BWire File.
74. For a short description of the project, see *About the Open Directory Project* (visited Feb. 22, 2000) <http://dmoz.org/about.html>.

engines.[75] AltaVista displays Open Directory results at the end of regular search results, HotBot uses some Open Directory data, and Lycos uses the Open Directory exclusively for its directory results.[76] Interestingly, Google has manipulated the Open Directory into a new form. Google has indexed all the Open Directory sites, applied its PageRank technology to them, and made its reranked version of the Open Directory. Initial searches of Google's version of the Open Directory indicate that it performs extremely well in finding law-related sites.

¶90 Even if search engines share a directory, they do not share the same search algorithm or the same index. Consequently, using the same search on two search engines that share a directory will yield different result sets. For example, the same search on HotBot and LookSmart does not produce the same results even though HotBot uses the LookSmart directory. This is because LookSmart emphasizes results from its directory on search results, but HotBot does not. LookSmart users are first presented with directory categories, then specific Web sites from within its directory, and finally Web sites from its index. HotBot, on the other hand, presents only one or two categories, then Direct Hit results, and finally Web pages from the Inktomi index.

¶91 Well-crafted directories, with quality Web sites correctly categorized, are a useful tool for any Web user, but they are particularly beneficial to legal researchers. Legal researchers tend to use sites that are usually selected for inclusion in directories and properly categorized therein. In the known item retrieval study described earlier, Yahoo! did better than all other search engines. Its excellent directory of Web sites was probably a factor in this. Consequently, legal researchers may want to favor search engines that offer human-compiled directory results at the top of result lists, specifically Yahoo!, and any engines using the Open Directory or LookSmart's directory.

Web Not Fully Indexed

¶92 There have been several recent studies to determine the approximate size of the Web, its characteristics, and the percentage of it indexed by search engines. Most come to similarly depressing conclusions: the Web is incomprehensibly large and getting larger, parts of it are inherently unknowable, and search engines are covering a diminishing percentage of it in their indexes. The following is a brief survey of recent Web studies to provide some background on the changing size and characteristics of the Web and its coverage by search engines.

---

75. Netscape purchased the project in 1998 but still allows other search engines to use the data. *See* Chris Sherman, *Humans Do It Better: Inside the Open Directory Project,* ONLINE, July 1, 2000, at 43, 46. In January 2000, the project had 21,500 contributors, 1,400,000 sites categorized, and 200,000 categories. *See* Elizabeth Weise, *Web Changes Direction to People Skills*; *Neatly Categorized Information Requires the Human Touch,* USA TODAY, Jan. 24, 2000, at 1D.

76. *See Open Directory Project Sites Using ODP Data* (visited Feb. 22, 2000) <http://dmoz.org/Computers/Internet/WWW/Searching_the_Web/Directories/Open_Directory_Project/Sites_Using_ODP_Data/>.

Lawrence and Giles Studies

¶93 Steve Lawrence and Lee Giles of the NEC Research Institute reported in July 1999 that the search engine with the largest index, Northern Light, indexed only about 16 percent of the publicly indexable Web.[77] This was a significant decrease from their original 1998 study, which concluded that 32 percent of the publicly indexable Web was indexed.[78]

¶94 In the 1998 study, Lawrence and Giles used 575 queries on six engines: HotBot, AltaVista, Northern Light, Excite, Infoseek, and Lycos. Queries were collected from real searches done by scientists at NEC. Making certain assessments about the independence of any search engine index, the size of the Web was estimated by analyzing the overlap of results from search engines. In December 1997, they estimated the size of the indexable Web to be between 90 million to 320 million pages. Using the more likely 320 million figure, the search engines varied in the percentage of the Web they indexed from HotBot at 34 percent to Lycos at 3 percent. With fairly low overlap, the six search engines collectively covered about 60 percent of the Web.[79]

¶95 In their follow-up study in February 1999, they expanded their study to include eleven engines and 1,050 queries. Web size estimation was done differently this time. Rather than analyzing the overlap of results from engines, they crawled all the pages on 2,500 random servers to determine a mean number of pages per server (289). They then approximated the number of Web servers to 16 million in total, of which 2.8 million were on the publicly indexable Web. Using the mean number of pages per server, and the approximate number of servers on the public Web, they estimated the indexable Web to be 800 million pages.[80]

¶96 Frighteningly, none of the engines indexed more than a sixth of the Web. Together the engines only covered about 42 million of the estimated 800 million pages, and overlap of the engines was low. Freshness of the indexes was also low; new or changed pages could take up to six months to be included in indexes. The percentage of invalid links ranged from 9.8 percent to 2.2 percent.[81] This time,

77. Lawrence & Giles, *Accessibility of Information*, *supra* note 5, at 108. They do not clearly define "publicly indexable," but it seems to mean nonpublic Web sites such as password-protected sites and sites containing dynamic content.

78. Lawrence & Giles, *Searching the World Wide Web*, *supra* note 5, at 99.

79. *Id*. at 99–100.

80. Lawrence & Giles, *Accessibility of Information*, *supra* note 5, at 107. In the past year this figure has significantly increased. One study published July 10, 2000, by an Internet company called Cyveillance found that 2.1 billion unique, publicly available pages exist on the Internet and that it is growing by 7 million pages per day. *See* Alvin Moore & Brian Murray, *Sizing the Internet* (visited Aug. 1, 2000) <http://www.cyveillance.com/newsroom/3012.asp>.

81. This correlates with a more recent study that determined over 10% of Web links are broken. *See* Mark Ward, *Web Links that Stick,* BBC NEWS ONLINE (June 14, 2000) <http://news.bbc.co.uk/hi/english/sci/tech/newsid_790000/790685.stm>.

individual search engines varied in coverage of the Web from Northern Light at 16 percent to Euroseek at 2 percent.[82]

¶97 For legal researchers, this may or may not be alarming. Subject areas were not analyzed in determining the percentage of the Web indexed. That is, the research did not specify the percentage indexed by subject area (i.e., of all government information, X percent is indexed). It is entirely possible that Web content in certain subject areas such as law, government, education, and medicine are disproportionately indexed. This seems to be true for legal information based on the known item study described earlier.

Web Connections Study

¶98 IBM, Compaq, and AltaVista recently completed a large-scale study analyzing the size and connections of the Web. They determined that, when Web connections are graphed, the shape of the Web is like a bow tie, a center circle with side triangles. The center circle represents a strongly interconnected core consisting of about 30 percent of all Web pages. The two sides of the bow tie are comprised of about 24 percent of all Web pages each. One side links into the core, but the core does not link back to it. The other side does not link to the core, but the core links to it. Finally, about 22 percent of Web pages are utterly isolated, or isolated in tiny clusters.[83] This conclusion completely negated a more optimistic study published in September 1999 concluding that any Web page was, on average, merely nineteen clicks away from any other page on the Web.[84]

¶99 If large portions of the Web are not connected, this implicates the potential of crawlers to capture a majority of the Web into an index. New approaches will have to be developed to help crawlers reach the isolated portions of the Web.

Difficulty of Determining Search Engine Index Size and Web Size

¶100 Besides the occasional results from such in-depth studies, users must rely on self-reporting by search engine companies about their index size. Although the statistics are probably reliable, self-reporting is somewhat problematic. Search engine companies are motivated to attract advertisers by claiming large indexes and high usage statistics.

¶101 There are two excellent Web sites with substantive information about search engines that is difficult to find elsewhere, including self-reported index

---

82. Lawrence & Giles, *Accessibility of Information, supra* note 5, at 108.
83. The study was published on the Web in May 2000. Andrei Broder et al., *Graph Structure in the Web* (visited July 31, 2000) <http://www.almaden.ibm.com/cs/k53/www9.final/>. For a simpler explanation of the study's conclusions, see Ian Austen, *Study Reveals Web as Loosely Woven,* N.Y. TIMES, May 18, 2000, at G8.
84. *See* Reka Albert et al., *Internet: Diameter of the World-Wide Web,* NATURE, Sept. 9, 1999, at 130, 130.

size. The two sites are Search Engine Watch by Danny Sullivan[85] and Search Engine Showdown by Greg R. Notess.[86]

¶102 Since the revelation about the low percentages of the Web covered by search engines, there has been a phenomenon sometimes referred to as "size wars." Search engine companies have engaged in an aggressive race to achieve (and boast about) the largest index.[87] It remains to be seen whether or not these new, gargantuan indexes can be kept current and free of dead links.

¶103 Legal researchers should keep apprised of Web size assessments and the competing index sizes of search engines. Arguably, however, ranking mechanisms are far more important than index size. During the known item study, for instance, Northern Light had the largest index of the search engines included in the study. It performed only average in known item retrieval, ninth out of fifteen. HotBot had the smallest index, yet performed better than all other engines except Yahoo!. This seems to indicate that index size is not necessarily related to search engine performance. Even so, search engines with the largest indexes will probably be the most likely to retrieve obscure items.

Blind Spots and the Invisible Web

¶104 Depressingly, no matter how large an engine's index, there are certain types of Web documents that it will never include. The blind spots of search engines vary somewhat by engine, but in general they include: pages with low link popularity, information in images and frames, dynamic content, and pages on password-protected sites. Most pages that fall into these blind spots are not indexed and, consequently, are not included in result lists from engines.

¶105 Unless many pages point to a page, engines will often miss that page or take a long time to crawl and index it.[88] Engines' indexes are built by "crawlers" (also called robots or spiders), pieces of software that start from a random page and follow links to find new pages to index, or old pages to re-index. The fewer pages that link to a page, the less likely it is that a crawler will hit the page, and the less likely the page will be added to a search engine index. Although added

---

85. Sullivan, *supra* note 6. For index sizes, see Danny Sullivan, *Search Engine Sizes* (last modified July 7, 2000) < http://www.searchenginewatch.com/reports/sizes.html >.
86. Notess, *supra* note 6. For index sizes, see Greg R. Notess, *Search Engine Statistics: Database Total Size Estimate* (last modified July 6–7, 2000) <http://www.notess.com/search/stats/sizeest.shtml>.
87. Notess's Search Engine Showdown contains an extremely useful "Search Engine Statistics" section describing search engine total index sizes. It contains a current chart as well as links to previous index sizes for the past year. Comparing statistics from January 1999 to July 2000, the smallest index (HotBot) has increased from about 20 to 280 million pages. The largest index (Northern Light) has increased from about 115 million to 282 million pages. Now one of the largest indexes, Google boasts that its index contains 560 million pages. Greg Notess, *Search Engine Statistics: Database Total Size Estimates* (visited July 31, 2000) <http://www.searchengineshowdown.com/stats/sizeest.shtml>.
88. *See* Stephen Manes, *Personal Computers; Why Web Search Engines May Speed Past Missing Links,* N.Y. TIMES, Feb. 11, 1997, at C3.

slowly, pages with low link popularity will eventually be included in search engine indexes.

¶106 Another blind spot for search engines are pages within password-protected sites. Crawlers do not have the capacity to enter password-required sites, even if those sites only require a free registration for a password. Hence, search engines do not crawl and index many substantive, free sources on the Web, such as newspaper sites like the *New York Times*.

¶107 Engines have trouble indexing information in frames. Certain engines can crawl them and deliver the information in a result list,[89] but when called up, the information does not appear in the context of its framed environment. There are some products that can index, and link correctly to, frame-based sites,[90] but not all search engines license or use such technology.

¶108 Less a blind spot than a willing blind eye, search engine crawlers turn away from sites on which they encounter the Robots Exclusion Protocol. The Robots Exclusion Protocol allows Web site managers to prevent parts of their site, or their entire site, from being indexed by search engines. They generally use it because they change site content frequently, are extremely security conscious, or worry about powerful crawlers bogging down their Web server.[91]

¶109 Search engines will not retrieve information contained in graphics. Although alt tags describing a graphic can be indexed, words in a graphic file cannot. Also, text in images such as text scanned and stored in .gif or .jpg files, or text within .pdf files, is invisible to search engines. Nontext images such as graphs, medical pictures, art reproductions, etc. are also not indexed.[92] Interestingly, there are experimental search engines in development which will allow users to search image databases using an image as a query, but these are not yet a public reality.[93]

---

89. AltaVista, Northern Light, and Google are capable of reading frames. *See id*.
90. MondoSearch is an example of a search software that can index frame-based sites properly. *See MondoSoft, Web Site Search Expert Receives 7.5M USD Investor Funding,* PR Newswire, Oct. 25, 1999, *available in* LEXIS, News Library, PR News File.
91. *See* Ken Phillips, *Untangling the Web; Internet Spiders,* PC WEEK, Apr. 1, 1996, at N20, N21.
92. There are specialty image-searching services available through some general engines. These services search a separate index of thumbnail versions of graphics from Web pages. Queried terms must appear in the name of the graphic file or in nearby text to be a hit. For example, AV Photofinder (http://image.altavista.com), Image Surfer (http://ipix.yahoo.com), and Scour.net (http://scour.net) all work this way. HotBot and some other engines allow users to specify that pages in result lists must contain images. This does not mean that the images themselves are indexed. It merely means the result list will contain pages with a searched term in the page's text and have a graphic file of some kind on the page.
93. BlobWorld, still in development, will allow users to compose an image query by submitting an image, viewing a feature-extracted "BlobWorld" version of the image, selecting the blobs they want matched, and specifying the importance of blob features. Queries are run against the database/index to find like images. IBM is also developing an image database system. Called Query By Image Content (QBIC), it allows users to specify properties of a desired image. *See* Ray R. Larsen & Chad Carson, *Information Access for a Digital Library: Cheshire II and the Berkeley Environmental Digital Library, in* KNOWLEDGE: CREATION, ORGANIZATION AND USE, PROCEEDINGS OF THE 62ND ANNUAL MEETING OF THE AMERICAN SOCIETY FOR INFORMATION SCIENCE 515, 521 (Larry Woods ed. 1999).

¶110 The largest search engine blind spot by far is dynamic content, often referred to as the invisible Web. Dynamic content is information that is served from databases through the Web in response to user queries. Contents of databases available through the Web are not actually *on* the Web in static form and are not indexed. This dynamic content, or nonstatic content, is invisible to search engines. (A telltale sign of pages with dynamic content is a URL containing a question mark or ending with an .asp, .cfm, .cgi, or .shtml extension.) Unfortunately, this part of the invisible Web is growing.[94]

¶111 Of all search engine blind spots, dynamic content is the biggest problem for *legal* researchers. Huge amounts of information from legal sites and government sites are being served from databases.[95] Educational institutions,[96] including law schools, are increasingly mounting dynamic content[97] as are governmental agencies. Some sites with judicial opinions are even using dynamic content.[98] Even simple directory information such as membership listings on the American Association of Law Libraries Web site or the home page of the U.S. Congress are now served from a database rather than existing in static html form. This is not to say that Web front-ending useful legal databases is a bad thing for the legal community. It is simply an unfortunate side effect that search engines cannot be used to locate the information contained in them.

¶112 Software tools are being developed to address this problem. One example called LexiBot, proprietary software available from BrightPlanet, supposedly searches both the static and dynamic content of the Web. Its creators recently completed a study in which they concluded that the size of the Web with dynamic content is five hundred times larger than the static, or publicly indexable Web.[99] In a second example, one that is publicly available, larger, and more general in scope, Lycos and Intelliseek have produced InvisibleWeb.com (http://InvisibleWeb.com), a searchable list of about ten thousand databases invisible to search engines. Lycos provides links to relevant databases from InvisibleWeb.com

---

94. *See* Whit Andrews, *Challenge For Spiders; Searching Invisible Web,* WEB WK., Feb. 3, 1997, at 48.
95. InvisibleWeb.com lists 1193 databases under the category of government databases and 1284 under the category of legal databases. *InvisibleWeb.com, The Search Engine of Search Engines* (visited Nov. 12, 1999) <http://invisibleWeb.com>.
96. Libraries are increasingly "front ending" databases onto the Web. As technologies facilitating this have become easier to use, more dynamic content is being mounted. *See* Kristen Antelman, *Getting Out of the HTML Business: The Database-Driven Web Site Solution,* 18 INFO. TECH. & LIBR. 176, 180 (1999).
97. In November 1999, the Center for Computer Assisted Legal Instruction (CALI) conducted a survey of dynamic Web content in law school Web sites. Of the 32 institutions that responded, 26 had dynamic content on their Web site. *CALI Technology Survey: Dynamic Web Content in Law School Websites* (Nov. 2, 1999) <http://www.cali.org/survey/1999/DynamicWebContent.html>.
98. For instance, U.S. 9th Circuit Court opinions are delivered from a database and do not exist in static html form. As a result, it is not possible to use a search engine to retrieve these opinions using keywords from the case as a query. Fortunately, a static page listing the names of the cases and docket numbers has been created. This static page can be captured by search engines.
99. *See Bright Planet Unveils the "Deep Web:" 500 Times Larger than Existing Web* (visited July 31, 2000) <http://brightplanet.com/newsroom/deepWeb.asp> (provides link to the study white paper.)

in search results.[100] In the meantime, legal researchers relying on search engines to find information on the Web may miss valuable material.

Commercial Nature of Search Engines

¶113 Search engine manipulation by Web developers and the commercialism of search engine companies is less problematic than the lack of coverage of search engines. Thankfully, both of these phenomena pose only slight problems for legal researchers using the Web.

¶114 Some search engines sell placement in result lists to the highest bidder. This practice is called pay for display, pay for performance, or pay for placement. Luckily, the developers of Web sites useful to legal researchers are unlikely to bid on search terms in order to ensure that their pages show up on result lists.

¶115 The largest of the search engines using pay for display is GoTo.com. GoTo.com allows companies to bid on search terms. When those terms are used in queries, the page of the company with the winning bid will be pushed to the top of the result list. That company then pays GoTo.com a predetermined amount, from a few cents to a few dollars, each time a user clicks on its link.[101] GoTo.com result lists differentiate between paid placements and other results by showing the amount it costs the company when users click on a link. GoTo.com claims to have strict relevancy requirements; for instance, you cannot bid on a term unless it is related to your company or association. Launched in June 1998, as of April 2000, it had approximately 25,000 advertisers.[102]

¶116 Although this sounds like an alarming development, legal researchers are not heavily affected by this practice. In the course of the known item retrieval study described earlier, fifty searches for law-related pages were completed. No paid placements appeared in the top twenty hits for *any* of the fifty searches on GoTo.com. This is most likely due to the fact that the types of pages that contain legal information are not usually hosted by companies willing, or needing, to pay for placement in result lists.[103]

---

100. *See* Paula J. Hane, *IntelliSeek Launches InvisibleWeb.com,* INFO. TODAY, Dec. 1999, at 32, 32; *Lycos Aligns With Intelliseek to Reveal the Invisible Web Catalog—The Largest Collection of Searchable Databases,* Bus. Wire, June 7, 1999, *available in* LEXIS, News Library, BWire File.

101. Although GoTo.com is unabashedly commercial, it was included in the known item and relevancy studies earlier in this article. It met selection criteria and has one of the higher user retention rates for search engines. *See GoTo.com Vaults 64 Percent to 24th Most Trafficked Web Site with 4.2 Million Unique Visitors,* Bus. Wire, Nov. 19, 1998, *available in* LEXIS, News Library, BWire File.

102. *See* Karen Kaplan, *The Cutting Edge: Focus on Technology Pay-Per-Click Concept Gets GoTo.com Farther,* L.A. TIMES, Apr. 3, 2000, at C1. Besides GoTo.com, only about 30 other search engines currently use pay for display, and these are among the less well-known search engines. *See PayPerClickSearchEngines.com* (visited July 24, 2000) <http://www.PayPerClickSearchEngines.com>.

103. The caveat is that law firms and lawyer referral services do participate in GoTo.com's term selling scheme. Entering the term "lawyers" generates a result list replete with paid-for placements. In a search I conducted on July 23, 2000, the Law Offices of Mark Siegel was the top bidder on "lawyers," having agreed to pay $3.25 for each user click through from GoTo.com results.

Spamdexing

¶117 A combination of the terms spamming and indexing, spamdexing is the practice of some Web developers who intentionally fool search engines into putting their pages undeservedly high on result lists.[104] Spamdexing is accomplished by many methods, including: creating multiple title tags, keyword stuffing on meta tags, keyword stuffing in html comments, keyword stuffing using text the same color as the background of the page or in tiny font, registering multiple duplicate pages, running software that creates "click streams,"[105] and creating dummy pages with nothing but multiple links to a page.[106] These behaviors exploit the relevancy-ranking algorithms of search engines that use title tags, keyword tags, position and frequency of search terms on a page, and the ratio of search terms to all other terms on a page to determine which pages to place at the top of result lists.

¶118 Developers of commercial sites spamdex because they want their site to appear often and high on search engine result lists. For example, realizing that many Web searches are for pictures of Pamela Lee Anderson, a Web developer might put that name one hundred times or so in the keyword tags of a Web page, even if the Web page is completely unrelated to Pamela Lee Anderson. When a user enters "Pamela Lee Anderson" into a search engine, the ranking algorithms of some search engines would force that irrelevant page to the top of the result list.

¶119 Thankfully, the sites most relied upon by legal researchers are not created by developers with commercial motivations. In general, developers of sites with legal information such as law school sites, court sites, government sites, etc., have little or no motivation to spamdex. For instance, the Legal Information Institute at Cornell would be unlikely to load its pages with hidden occurrences of "Britney Spears" or "Pokemon." Moreover, those motivated to spamdex are not likely to use law-related terms to force search engines to deliver their sites undeservedly in result lists. For instance, a commercial company is not likely to stuff keywords like "Thomas, Library of Congress" on its page to garner placement in result lists. Consequently, although spamdexing is the bane of many a search engine developer's existence, it does not heavily impact legal researchers using the Web.

---

104. For an excellent summary of spamdexing practices and their legal implications, see Ira S. Nathenson, *Internet Infoglut and Invisible Ink: Spamdexing Search Engines With Meta Tags,* 12 HARV. J.L. & TECH. 43 (1998).

105. Scripts can be created that automatically run a search on a search engine, select a certain page from the result list, linger at that page for a few moments, and then begin again. This inhuman "click stream" spamdexes engines that rank pages based on their popularity with previous search engine users, such as Direct Hit.

106. This arguably spamdexes engines that rank pages based on the number of times they are linked to by other pages, such as Google and other search engines using link analysis in their ranking algorithms.

Conclusion

¶120 With the increasing amount of legal information available on the Web, search engines have become a vital legal research tool. Bookmarks, legal directory sites, and memorized URLs can no longer suffice to tame the morass of Web pages relevant to law. Consequently, it is incumbent upon legal researchers to critically evaluate the performance of search engines in locating legal information.

¶121 We must determine effective methods to compare Web search engine performance just as we have developed methods for comparing the performance of other online legal resources. Until that point is reached, researchers are left with the anecdotal conclusions presented here regarding Web search engine performance for law-related searching. Specifically, these studies indicate that general Web search engines outperform legal search engines in retrieving known items. Legal search engines, however, outperform general search engines in retrieving items relevant to legal topical queries. Finally, the studies indicate that legal information on the Web is well indexed by search engines and that new ranking methodologies greatly benefit legal researchers.

Appendix A
Search Engines Used

AltaVista	http: //www.altavista.com
AltaVista's LawRunner	http: //www.lawrunner.com
Excite	http: //www.excite.com
Fast	http: //www.alltheweb.com
Findlaw's LawCrawler	http: //lawcrawler.findlaw.com
Go Network's Infoseek	http: //infoseek.go.com
Google	http: //www.google.com
GoTo.com	http: //www.goto.com
HotBot	http: //www.hotbot.com
LookSmart	http: //www.looksmart.com
Lycos	http: //www.lycos.com
Northern Light	http: //www.northernlight.com
Snap	http: //www.snap.com
WebCrawler	http: //www.webcrawler.com
Yahoo!	http: //www.yahoo.com

Appendix B
Known Item Searches (in Order of Date Performed)

Known Item
URLS Counted as Correct Item
Date of Search and Search Used

1. GAO Report: AIMD-98-155, Air Traffic Control: Weak Computer Security Practices Jeopardize Flight Safety
   www.gao.gov/AIndexFY98/abstracts/ai98155.htm
   www.fas.org/irp/gao/aimd-98-155.htm
   www.bts.gov/ntl/data/ai98155.pdf
   9/9/1999; gao report and "air traffic control weak"

2. Alaska Law Review 15: 33-58, Sidestepping Scott: Modifying Criminal Discovery in Alaska by Cameron J. Williams, June 1998
   www.law.duke.edu/journals/alr/alrtoc15n1.htm
   www.law.duke.edu/shell/cite.pl?15+Alaska+L.+Rev.+33
   9/9/1999; "alaska law review" and sidestepping

3. Department of Defense Procurement Statistics: Prime Contract Awards Fiscal Year 1998
   http://web1.whs.osd.mil/peidhome/procstat/p03/fy1998/p03.htm
   http://web1.whs.osd.mil/peidhome/procstat/procstat.htm
   http://web1.whs.osd.mil/peidhome/prodserv/fy1998/p07.htm
   http://web1.whs.osd.mil/peidhome/geostats/p09/fy1998/p09.htm
   9/10/1999; "department of defense" and "prime contract awards" and 1998

4. Center for Reproductive Law and Policy Report "Adolescent Reproductive Rights," Feb. 1999
   www.crlp.org/icpdadolescents2.html
   www.crlp.org/articles_pubsub.html
   www.crlp.org/0399jbspeech.html
   9/10/1999; "adolescent reproductive rights" and "center for reproductive law"

5. Bare v. Liberty Mutual Fire Ins. Co., Montana Supreme Court, May 1998, Docket # 97-434
   www.lawlibrary.mt.gov/97-434.htm
   9/10/1999; "bare v liberty mutual fire" and "workers compensation"

6. Environmental Protection Agency, Office of Research and Development, "Oxygenates in Water: Critical Information and Research Needs"
   www.epa.gov/ncea/oxyneeds.htm
   9/11/99; "oxygenates in water" and epa

7. Journal of Land Use and Environmental Law, 11: 325-374 The Best Laid Plans: The Rise and Fall of Growth Management in Florida by Mary Dawson, Fall 1996
   www.law.fsu.edu/journals/landuse/vol112/dawson.html
   www.law.fsu.edu/journals/landuse/112.html
   9/11/1999; "best laid plans" and florida and "journal of land use"

8. Executive Order 13031, 12/13/1996, "Federal Alternative Fueled Vehicle Leadership"
   http://policyworks.gov/org/main/mt/homepage/mtv/eo13031.htm
   www.epa.gov/docs/fedrgstr/EPA-AIR/1996/December/Day-17/pr-23907.html
   http://pub.fss.gsa.gov/environ/eo/eo13031.html
   www.afdc.nrel.gov/documents/execord96.html
   www.pub.whitehouse.gov/uri-res/I2R?urn:pdi://oma.eop.gov.us/1996/12/14/3.text.1
   9/11/1999; "executive order" and "alternative fueled vehicle leadership"

9. Hughes Aircraft Co. v. Jacobson, U.S. Supreme Court, Jan. 1999, Docket # 97-1287
   http://supct.law.cornell.edu/supct/html/97-1287.ZO.html
   www.law.emory.edu/fedcircuit/june97/94-5149o.html
   http://usscplus.com/current/cases/ASCII/9800025.txt
   9/11/1999; "hughes aircraft co v jacobson" and "employee benefits"

10. SEC Final Rule, 11/3/1998, OTC Derivatives Dealers; Correction Release No. 34-40594A
    www.sec.gov/rules/finrindx.htm
    http://ftp.sec.gov/rules/final/34-440594a.htm
    http://edgar.sec.gov/rules/final/34-40594a.htm
    9/11/1999; "release no 34-40594a"

11. Haddle v. Garrison, U.S. Supreme Court, 12/14/1998, Docket # 97-1472
    http://supct.law.cornell.edu/supct/html/97-1472.ZO.html
    (many other URLs counted)
    9/11/1999; "haddle v garrison" and "medicare fraud"

12. Public Law 105-42, "International Dolphin Conservation Program Act"
    http://thomas.loc.gov/cgi-bin//bdquery/z?d105:h.r.408:
    www.gpo.ucop.edu/gpo/catalog/97_presdoc18au97a.html
    (Counts as within two clicks but not as hit.)
    9/12/1999; "105-42" and "international dolphin conservation program act"

13. Splude v. Apfel, First Circuit, 1/19/1999, Docket # 98-1630
    www.law.emory.edu/1circuit/jan99/98-1630.01a.html
    9/12/1999; "splude v apfel"

14. National Conference of Commissioners on Uniform State Laws
    www.nccusl.org
    9/12/1999; "national conference of commissioners" and "uniform state laws"

15. Association of Southeast Asian Nations
    www.asean.or.id
    www.aseansec.org
    9/12/1999; "association of southeast asian nations"

16. National Conference of State Legislatures
    www.ncsl.org
    http://ncsl.org
    9/12/1999; "national conference of state legislatures"

17. Cervantes v. Drayton Foods LLC, North Dakota Supreme Court, July 1998, Civil # 970388, 1998 ND 138
    www.court.state.nd.us/court/opinions/970388.htm
    9/13/1999; "cervantes v drayton foods"

18. National Bankruptcy Review Commission
    www.nbrc.gov
    9/13/1999; "national bankruptcy review commission"

19. Nebraska State Bar Association
    www.nebar.com
    9/14/1999; "nebraska state bar association"

20. National Association of Secretaries of State
    www.nass.org
    9/14/1999; "national association of secretaries of state"

21. National Center for Agricultural Law and Research and Information
    http://law.uark.edu/arklaw/aglaw/
    9/16/1999; "national center for agricultural law research"

22. Berkman Center for Internet and Society
    http://cyber.harvard.edu
    www.law.harvard.edu/Programs/center_law
    9/161999; "berkman center for internet and society"

23. European Court of Justice
    www.europa.eu.int/cj/index.htm
    http://cuna.eu.int/en/

www.europa.eu.int/cj/en/index.htm
9/16/1999; "european court of justice"

24. United States Court of Appeals for the Fifth Circuit
www.ca5.uscourts.gov
9/17/1999; "court of appeals" and "fifth circuit"

25. Michigan Supreme Court
www.SupremeCourt.state.mi.us
www.state.mi.us/jud
www.voyager.net/supremecourt/index.htm
9/17/1999; "michigan supreme court"

26. Nevada Attorney General
www.state.nv.us/ag
http://colorado.state.nv.us/ag
9/18/1999; "nevada attorney general"

27. Senator Diane Feinstein
www.senate.gov/~feinstein
9/18/1999; "senator diane feinstein"

28. John H. Jackson, Professor of Law Emeritus at Michigan Law School
www.law.michigan.edu/faculty/jackson.htm
http://141.161.67.230/faculty/vitas/jackson-j.html
www.dfait-maeci.gc.ca/english/foreignp/dfait/policy_papers/1997/01/refdoc1-e.htm
9/20/1999; "john h jackson" and law professor and michigan

29. State Senator Connie Stokes—Georgia
www.ganet.org/senate/getsen.cgi?passval=Stokes,43rd
http://www2.state.ga.us/legis/1997_98/house/gasm43.htm
http://www2.state.ga.us/Legis/1999_00/senate/gass43.htm
http://franklin.vote-smart.org/state/Georgia/legis/Upper/43/b025997.html
www.afaga.org/GASenate.htm#Stokes
9/20/1999; "connie stokes" and georgia

30. Illinois Supreme Court Rules
www.illinoisbar.org/CourtRules
9/21/1999; illinois and "court rules"

31. North Carolina General Statutes
www.ncga.state.nc.us/Statutes/Statutes.html
http://204.84.200.30/html1999/statutes/toc.html
http://ftp.ncga.state.nc.us/Statutes/Statutes.html
9/21/1999; "north carolina general statutes"

32. New Jersey Rules of Professional Conduct
    www.njlawnet.com/nj-rpc/index.html
    http://njlawnet.com/nj-rpc/index.html
    9/22/1999; "new jersey rules of professional conduct"

33. Rhode Island General Laws
    www.rilin.state.ri.us/Statutes/Statutes.html
    www.rilin.state.ri.us/gen_assembly/genmenu.html
    9/24/1999; "rhode island" and statutes

34. Ohio Department of Taxation
    www.state.oh.us/tax
    9/25/1999; "ohio department of taxation"

35. Trade and Development Agency
    www.tda.gov
    9/25/1999; "trade and development agency"

36. Federal Election Commission
    www.fec.gov
    9/25/1999; "federal election commission"

37. Congressional Budget Office
    www.cbo.gov
    9/25/1999; "congressional budget office"

38. Massachusetts Executive Office of Health and Human Services
    www.magnet.state.ma.us/eohhs/eohhs.htm
    9/25/1999; massachusetts and "health and human services"

39. Bracton's De Legibus
    http://bracton.law.cornell.edu/bracton/Common/index.html
    http://supct.law.cornell.edu/bracton/Common
    10/3/1999; "bracton de legibus"

40. The Third Branch, The Newsletter of the Federal Courts
    www.uscourts.gov/ttb
    10/3/1999; "third branch" and federal courts and newsletter

41. Cardozo Law Review
    www.yu.edu/cardozo/journals/cardlrev/
    10/3/1999; "cardozo law review"

42. Los Angeles Municipal Code
    http://Cityfolio.ci.la.ca.us/
    http://cityfolio.ci.la.ca.us/cgi-bin/om_isapi.dll?clientID=82560
    www.codesite.com/LAMC/MTOC.HTM (Incomplete, but will count.)
    10/3/1999; los angeles municipal code

43. Internal Revenue Bulletins
    www.irs.ustreas.gov/prod/ind_info/bullet.html
    www.irs.ustreas.gov/plain/bus_info/bullet.html
    www.fedworld.gov/ftp/irs-irbs/irs-irbs.htm
    www.fedworld.gov/pub/irs-irbs/irs-irbs.htm
    10/7/1999; "internal revenue bulletin" and "internal revenue service"

44. European Investment Bank
    www.eib.org
    www.eu.int
    www.bei.org
    10/8/1999; "european investment bank"

45. National Library of Canada
    www.nlc-bnc.ca/ehome.htm
    www.nlc-bnc.ca/
    http://nlc-bnc.ca
    10/8/1999; "national library of canada"

46. White House Millennium Council
    www.whitehouse.gov/Initiatives/Millennium/index.shtml
    10/8/1999; "white house millennium council"

47. Kluwer Law International
    www.kli.com/
    www.kluwerlaw.com and www.kluwerlaw.com/index.htm
    10/8/1999; kluwer law international

48. All Hands, Magazine of the U.S. Navy
    www.chinfo.navy.mil/navpalib/allhands/ah-top.html
    www.mediacen.navy.mil/pubs/allhands/contents.htm
    10/8/1999; "all hands" and navy

49. Missouri Revised Statutes
    www.moga.state.mo.us/homestat.htm
    www.moga.state.mo.us/statutes/statutes.htm
    www.senate.state.mo.us/statutes.htm
    10/8/1999; missouri statutes

50. Moriarty v. Svec, Seventh Circuit, Dec. 14, 1998, Docket # 98-1849
    www.kentlaw.edu/7circuit/1998/dec/98-1849.html
    10/10/1999; "moriarty v svec" and seventh circuit

Appendix C
Relevancy Searches in Order of Date Performed

Information Sought, Search Date, Search Used, Points Awarded Based on Information Retrieved.

Biographical information on Ernst Rabel
- Search date: Nov. 20, 1999
- Search: "ernst rabel"
- 3 points: paragraph or longer about him
- 2 points: at least two sentences about him
- 1 point: sentence about him, important dates, or lists of works by or about him
- 0 points: name appears only, dead link, or duplicate hit

Convention for the Conservation of Salmon in the North Atlantic Ocean and conservation of salmon in the North Atlantic generally
- Search date: Nov. 21, 1999
- Search: "convention for the conservation of salmon in the north atlantic"
- 3 points: full text of treaty
- 2 points: link to full text of treaty or paragraph summary or discussion of it
- 1 point: information about this treaty's status or summary of other salmon protection treaties
- 0 points: name of treaty only, dead link, or duplicate hit

National Commission on Reform of Federal Criminal Laws
- Search date: Nov. 27, 1999
- Search: "national commission on reform of federal criminal laws"
- 3 points: more than a paragraph about it, or its final report
- 2 points: at least a paragraph about it, or a link to its final report
- 1 point: citations of its works, or a sentence about it
- 0 points: name of commission only, false hit, duplicate hit, or dead link

Discovery rules for electronic evidence
- Search date: Nov. 28, 1999
- Search: discovery electronic evidence
- 3 points: full-text paper on it (not by a vendor), gateway site with active links
- 2 points: more than a paragraph of discussion about it, or a vendor site with at least three paragraphs of information
- 1 point: a few sentences about it, or a vendor page with more than one paragraph about it
- 0 point: phrase only, false hit, duplicate hit, or dead link

New York State Commission on Judicial Conduct
Search date: Nov. 29, 1999
Search: new york state commission on judicial conduct
3 points: its home page, or more than two-page paper about it
2 points: paragraph or more about it, link to its home page, or an internal page of its site
1 point: contact information, at least two sentences about it, or cites of its publications
0 points: false hit, duplicate hit, dead link, or name of council only

UCLA Bulletin of Law and Technology
Search date: Dec. 27, 1999
Search: "ucla bulletin of law and technology"
3 points: home page
2 points: link to home page
1 point: two sentences or more about it, contact information, or manuscript submission information
0 point: dead link, or mentioned only

Lawyer referral services for Chicago, and information about legal aid or legal clinics in Chicago or Illinois generally
Search date: Dec. 27, 1999
Search: lawyer referral service chicago
3 points: home page of lawyer referral service done by org. or gov. for Chicago or Illinois, or a list of legal services with contact information
2 points: link to a lawyer referral service home page, legal directories for Illinois or Chicago, or contact information for one service
1 point: lawyer referral nationally, or general information about hiring a lawyer
0 points: dead link, false hit, or duplicate hit

Appeals Council for the Social Security Administration (SSA), and information about the SSA adjudicative process
Search date: Dec. 28, 1999
Search: "appeals council" and "social security administration"
3 points: home page for the Office of Hearings and Appeals (OHA) at the SSA, more than three paragraphs about SSA adjudicative process, or full-text primary law about appeals proceedings in the SSA
2 points: link to SSA's OHA home page, at least two paragraphs about proceedings in the SSA, case or ruling about appeals council, or two-paragraph summary of case or statute about SSA appeals process
1 point: a few sentences about SSA courts, a brief summary of a case or rul-

ing mentioning SSA appeals council, or sentences about adjudicative processes in the SSA
0 points: dead link, duplicate hit, false hit, or mentioned on page

Licensing and Certification Programs for Midwives
Search date: Dec. 30, 1999
Search: midwifery certification licensing
3 points: home page of state licensing board, midwifery school home page with explanation of process and contact information, at least a page explaining licensing or certification programs, primary state law about midwivery, or home page of professional midwife organization
2 points: link to home page of licensing board, link to midwifery school home page, or lists of boards or schools in any state
1 point: one paragraph or less about licensing/certification of midwives or differences in scope of practice
0 points: false hit, dead link, duplicate hit, or words mentioned only

Freedom of religion clause in the Guatemalan constitution
Search date: Dec. 31, 1999
Search: guatemala constitution "freedom of religion"
3 points: more than three-paragraph discussion of freedom of religion in Guatemala, a copy of the Guatemalan constitution, or a site devoted to freedom of religion in Guatemala
2 points: link to Guatemalan constitution, at least one paragraph about freedom of religion in Guatemala, or a case about religious persecution in Guatemala
1 point: information about freedom of religion in Central America or Latin America
0 points: false hit, dead link, duplicate link, or mentioned only