LawBOT: an assistant for legal research

Sandip Debnath & Sandip Sen
Mathematical & Computer Sciences Department,
University of Tulsa
sandip-debnath@utulsa.edu, sandip@kolkata.mcs.utulsa.edu
Brent Blackstock
Attorney at Law, MoreLaw
blackstock@morelaw.com

Abstract

Intelligent agents are being deployed in diverse application domains. Both desktop based and Internet based personal assistant agents have been developed to assist users with their information processing chores [1, 2, 6, 5, 3, 7, 10]. In this paper we present an Internet based agent designed to assist legal researchers in retrieving laws and case reports electronically warehoused at a diverse set of databases maintained by local, state, and federal governments. LawBOT is implemented as a collection of agents which are employed according to users' preferences to collect, filter, organize and recommend relevant case histories, state statutes or supreme court cases. Our goal is to create a system that can be effectively used not only by lawyers but also by the lay person to retrieve legal documents relevant to the issue that the user wants to research. The requirement of enabling research by the commoner required us to add a novel ontology-based search component into LawBOT. We have developed an ontology for some of the common law categories. This ontology is used to map colloquial terms to corresponding legal terminology. This feature enables the average user to perform a more effective and thorough search for relevant legal documents. The ontology also enables query enhancement to search by related words which can return a more comprehensive set of documents.

1 Introduction

Legal research is the process of finding and organizing a collection of legal documents consisting of Constitutional provisions, Statutes and Cases that bear on the facts in question. It provides the raw material for Legal Analysis by furnishing the basic information from which a logical conclusion about appropriate conduct can be drawn given a set of facts and circumstances. Legal research is an iterative process. A researcher engaged in the process will probably discover new avenues of legal inquiry from the retrieved documents. These new avenues may suggest other forms of legal analysis that lead to additional areas of legal research which, in turn, will probably identify additional legal documents.

Legal information consists of two classes of documents:

Laws: Laws are abstract statements of the rights, privileges, duties, permissions and prohibitions that apply to persons within a nation or state.

Opinion: An opinion is an application of one or more Laws to a specific pattern of facts. In the United States and in many other nations, laws have a three-tiered hierarchical structure. At the top of the structure is a Constitution, which is a meta-rule stating how subsidiary laws come into existence, are interpreted and enforced. Typically, subsidiary laws, known as statutes, are enacted by a legislative body (often called a Congress or Legislature) pursuant the method stated in the Constitution. Statutes are abstract statements of either permitted or prohibited conduct applicable to classes of facts and circumstances. These statutes are then applied to facts by Courts in the judicial system. The logic and analysis employed by the Court may be recorded in the form of a written opinion known as Case. The collection of Cases serve as guidance for all in forecasting what might be the outcome of a given legal dispute.

Historically, legal documents in the form of Constitutions, Statutes, and Cases have been and still are published in hard copy form and are collected in libraries. The collected documents are indexed extensively according to legal concepts. The traditional method of legal research requires that a researcher infer relevant concepts from a given set of facts. Using the inferred concepts and the pre-compiled index, a researcher can then identify legal documents pertinent to the set of facts. As a result, legal research is both time consuming and labor intensive. The effectiveness of this process is limited by the accuracy of the indices, and the degree to which the legal researcher is able to infer what concepts are relevant to the set of facts he or she is analyzing.

With the advent of word processing, ever more documents, including legal documents have been stored electronically. Several private companies, including Westlaw [19], Lexis [13] and LOIS (Law Office Information Systems) [14] have collected or recreated extensive databases of legal documents. They charge fees based on connection time or a flat rate to access search engines connected to the databases. In addition to adding the convenience of immediate access to larger numbers documents than could be reasonably stored in a private library, the search engines provide the novel functionality of full text search capabilities. A search engine makes a full text search of a document feasible. The index generated by a search engine can have a much larger number of entries than a manual index, and therefore can contain a reference to every word in every document in the database. Full-text search has revolutionized the process of legal research by producing a paradigm shift in the manner in which legal research can be carried out. Rather than attempting to map a set of facts into the concepts in a manually generated index, the researcher may request a full-text search for certain words expected to appear in the document itself. Full text searching removes the need for an index of concepts and brings the process of legal research one step closer to the material being researched.

Three other factors are now reshaping the future of electronic legal research:

1. The cost of document storage has declined dramatically so that it is now much less expensive to store a document electronically than on paper.

2. The cost of accessing documents via a search engine has also declined with the widespread availability of powerful and yet inexpensive servers.

3. The Internet has made it possible to access documents cheaply and remotely.

The confluence of these three factors has generated a rapidly expanding global database of legal documents. Such applications will also be aided by advances in assisted browsing technology [12]. Consequently, governments have published an ever-growing proportion of their legal documents electronically, and have made them available on the Internet. Now it is possible for anyone with Internet access to search for and examine legal documents without having access to the hard copy and without having the legal training necessary to map patterns of facts into abstract legal concepts. One may, however, argue that the new research methods cannot be as effective as the traditional methods. Whether or not that is true, the fact remains that electronic research is in use, although it may not ever completely replace traditional means of legal research.

The development and proliferation of these Internet legal resources [8, 15] was uncoordinated. The syntax of search queries and the structure of the databases vary from domain to domain. A legal researcher who must access several different domains must learn how to use several different search engines. This situation provides for a convincing application for agent-based technology. Agents may be deployed to provide a consistent user interface, translate research requests into distributed queries, communicate via the Internet with the various search engines using the native syntax of each engine, and assemble, rate and order the appropriateness of the documents retrieved by multiple, distributed queries. Thus, the technical complexity of an extensive legal research may be hidden from the researcher, allowing him or her to focus attention on the logical analysis of the research itself. The system presented in this paper, LawBOT, was designed and implemented to provide this functionality and facilitate legal research over the Internet.

2 LawBOT Architecture and Implementation

The system architecture is presented in Figure 1. In the following we briefly state the functionality of each module:

Interface: The first screen is used to login to the system. User profiles including preferences and recent searches are stored in the system. These are used to present a customized second screen (e.g., allowing users to resubmit modified versions of recent searches, biasing search to look for specific states, etc.). The interface connects to the user preference database to collect or edit the user interactions and preferences. The interface forwards user queries to resource managers and also displays to the user the results provided by the resource manager.

Resource Manager: Resource Manager (RM) plays the important role of organizing information retrieval and processing. It augments the user query by consulting domain ontologies and then deploys the relevant Resource Agents as required by the query.

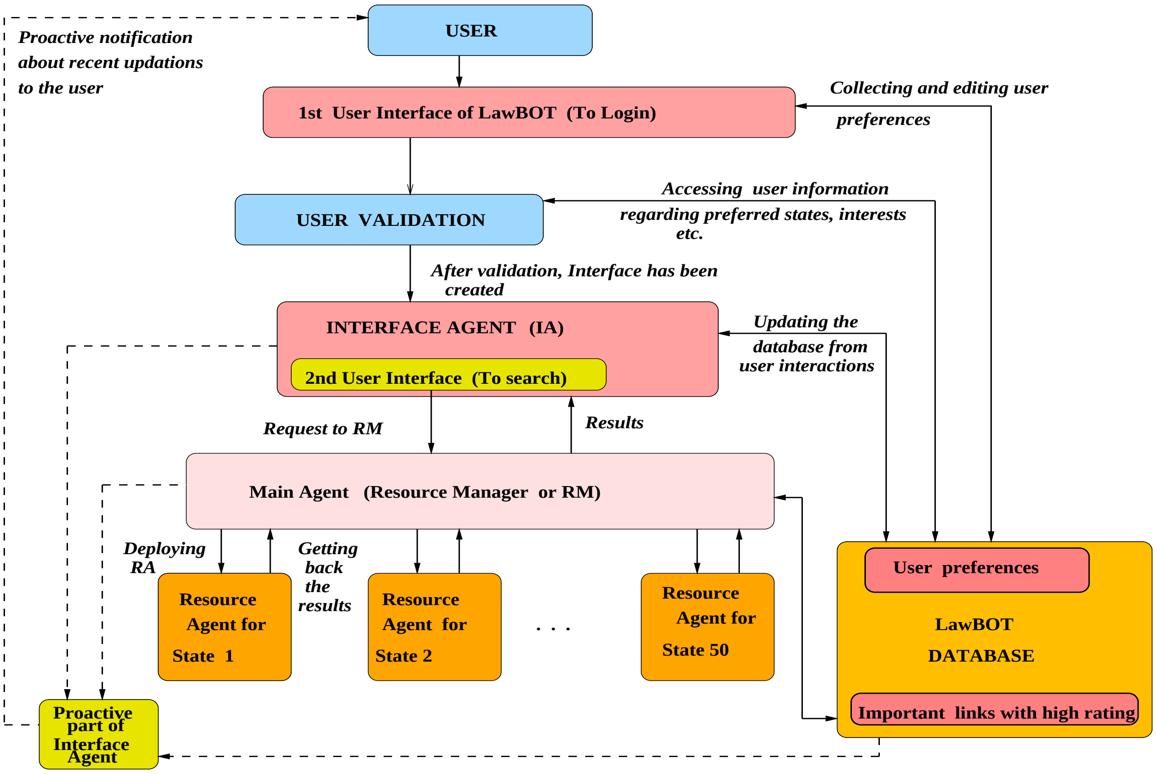

Figure 1: The LawBOT architecture.

The RM fuses the information retrieved by resource agents and from existing LawBOT database and returns this to the interface. The RM can also perform proactive notification to the users with recent updates on prior search results if requested. This feature allows the user to be updated with important and timely information as new laws and cases become available.

Resource Agent: The user's selection of jurisdictions to be searched determines which resource agent (RA) or agents will receive search requests. On receiving requests, the research agents format them according to the syntax expected by the search engine of the jurisdiction to be searched. The RAs are programmed to access specific sites on the web and run a word filtering algorithm over the retrieved documents. The documents are sorted and ordered according to word densities. The ordered documents are returned to the Resource Manager using an Agent Communication Language (ACL) [9, 11].

LawBOT Database: The LawBOT database stores user preferences, domain ontology, and auxiliary information like frequently used links, etc. using which the RM can effectively reformulate queries and rank results.

Proactive Component: The proactive component can use stored user preferences and past queries to do more extensive research off-line, i.e., when the user is not logged in. If new documents with high relevance is found in this off-line search, the user will be notified about the availability of new documents relevant to their interest.

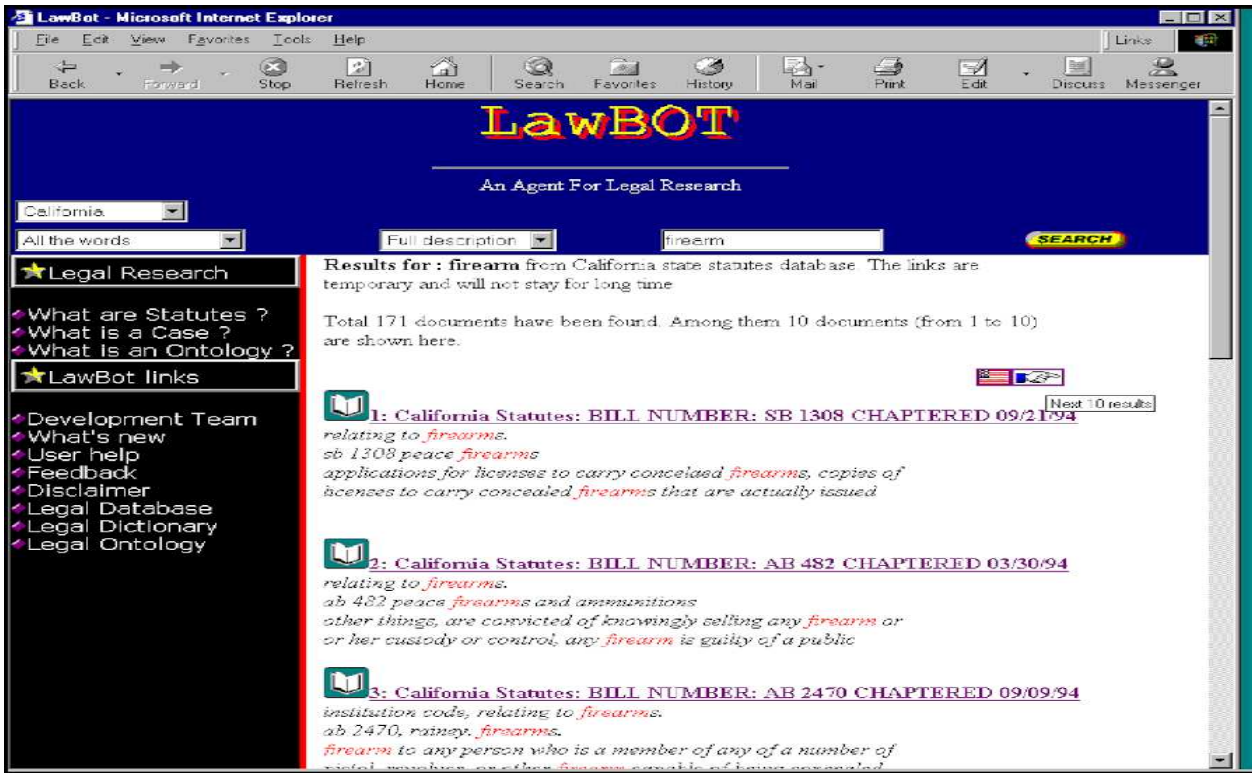

Figure 2: The result from searching the California statutes relating to firearm.

The user interface runs as a Servlet [17] application. The Resource Manager and Resource Agents are Java [16] based applications which keeps coordination among the Resource Manager.

Figure 2 shows the results generated by LawBOT for a search for statutes relating to *firearms* in the state of California. LawBOT displays the context of the query match for each statute retrieved. Based on our experience the availability of the context allows the user to quickly identify the statute of relevance. The user may examine any of the references returned by clicking it. Next to the reference is a box where the user may indicate a rating of the reference. The ratings are collected by LawBOT to assist other users that have issued the same search request. Although the usefulness of a reference depends on the particular requirements of each user, a higher rating by one user may suggest that the reference may be more useful to another user who issues the same search request.

3 Using ontology to augment search

As one of our principal goals for developing LawBOT was to enable the lay person to retrieve legal documents of interest, we had to provide additional functionality to rephrase an informal query into legal jargon. The use of an extensive ontology [18] was planned to enable this functionality. Ontology helps us build proper relations between words and phrases and those relations can be used to reformulate the query if very few or no results are available. It is very difficult for a common person to find the documents efficiently and quickly due to the lack of knowledge to use the legal terminology. Ontology gives the power to reformulate a particular phrase into a proper closely related legal jargon. The facility to reformulate a query helps naive users to find out legally useful documents from a huge database. We first discuss some examples to motivate the kind of functionality we planned for and then present the outline of our approach to obtaining such a capability.

Legally appropriate synonyms: The word *kid* is similar in meaning to *baby*, *child*, or *youngster*. Of these, *child* is used most frequently in law records. So any search with the words *kid* or *baby* will rarely produce any worthwhile result. LawBOT uses its legal ontology to search by *child* as well when a naive user chooses *kid* as the search word.

Related words and law categories: The word *child* may be associated with a number of other words, e.g., *care*, *custody*, *support*, *abuse*, etc. A user who searches only with *child* will receive too many references. We assist the user to rephrase such searches by asking the user to select either from a law category (e.g., family law, criminal law), or by expanding the search by choosing from a suggested list of often-used additional terms.

Appropriate search: Legal issues are categorized into standardized, fixed sections. The number is indicative of the category of the law. A search by *26* or *Section 26* should initiate the search for `Internal revenue` related laws. This is unique in the context of legal research. Such information is rarely used in the search engines which use either word count or word density. Rather, we often find common search engines returning irrelevant pages which include 26 in the date (April 26, 1996), or as in *Local 26* or in *26th Delaware Street*.

Our ontology is a semantic network (Figure 3) which relates words by their relationships. Words are associated with synonyms, hyponyms, hypernyms (this, for example, allows us to additionally search for firearm laws if someone asks for handgun laws), often-used associated words, special synonyms (e.g., *child* for *baby*), etc. If the user specified keywords contain a special synonym, the latter is used in place of the given keyword. If the number of documents returned for the user specified search is below a threshold, the search is augmented by using synonyms, hyponyms, and hypernyms. On the other hand, if the user specified search returns too many results, the user is asked to choose from law categories and/or augment search with often-used associated words. The current system uses a core ontology with terms from family law. This was appropriate to demonstrate the proof-of-concept of ontology based search augmentation, but needs to be expanded to other law categories.

4 How LawBOT Differs from Other Search Engines

LawBOT is different from other web agents or engines available on the Internet in the following ways:

- Unlike other general-purpose search engines, LawBOT is designed to facilitate legal research. Legal documents housed on a server are not typically indexed by MetaCrawlers [4]

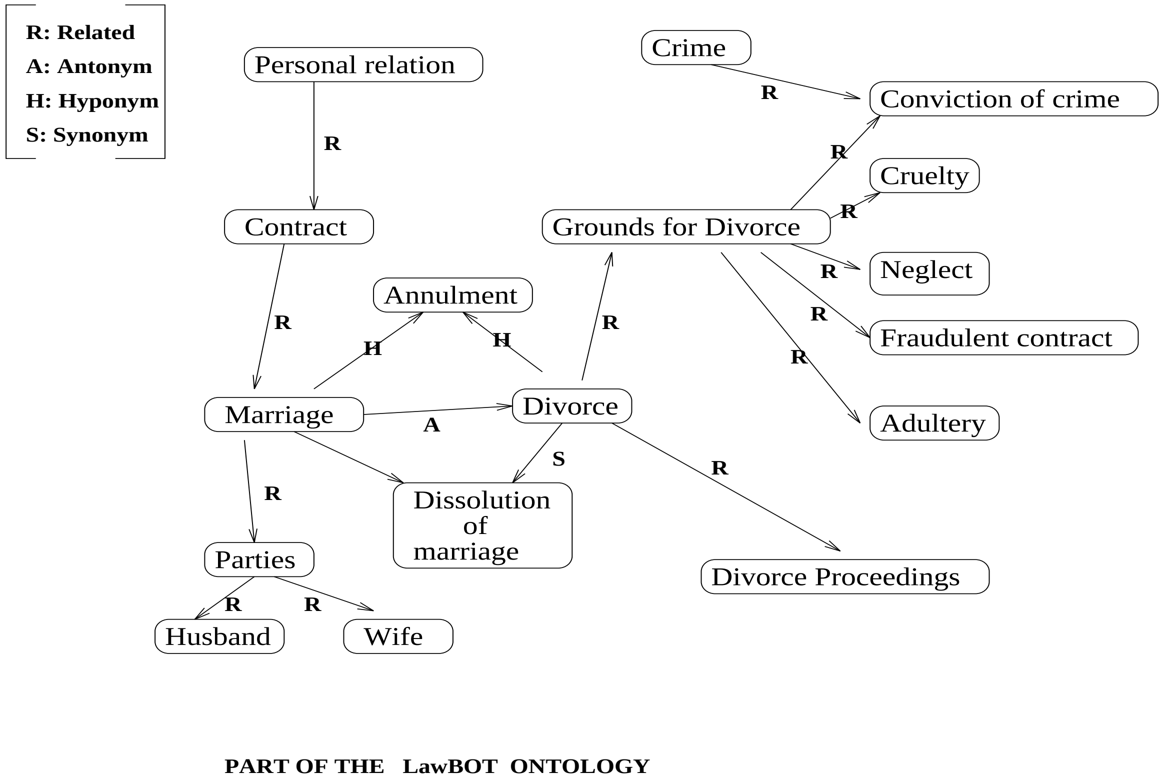

Figure 3: A small section of our ontology for family-law related words.

unless they also appear on a web page. Therefore, most legal documents are missed by a search request directed to a general-purpose search engine.

- The general search engines are nowadays keeping their own database of synonyms and sometimes a list of contextual words. But, as these are all general purpose search engines, it is almost impossible to know a particular user's context or intention for the search. The possible contexts of a given word or a set of words is much larger in general compared to a specific domain like legal search. Also, the likelihood of these different contexts will be vastly different depending on the particular domain of search. For example, the word "swimming" is more likely to be related to words like "gear", "equipment", etc when the particular domain is travel/leisure and is more likely to be related to words like "regulations", "accidents", etc. when the domain is legal research. It is very unlikely that the relevancy of documents returned by a well-designed domain specific search engine can be matched by a general purpose search engine.

- LawBOT allows the user to define the search domain so that documents retrieved by LawBOT originate only from official sources. Although it is possible to use LawBOT to access MetaCrawlers, the results of the search will probably include unofficial documents from sources other than the databases of laws and opinions.

- The systematic use of a well-formed ontology allows naive users to effectively search for legal documents from multiple sources.

- With each returned document link, we also present the context of the match by including the text around the match. We return only the first few match contexts. The tagging of returned references with the context of matching of the searched word(s) allows the user to quickly identify relevant documents.

5 Discussions

LawBOT is an implemented Internet based multiagent system for assisting in legal research. User specified keyword based queries are used to spawn multiple searches for legal documents (statutes, laws, cases) from local, state, and federal information repositories accessible through the web. A key feature of LawBOT is the ability to rephrase colloquial queries into corresponding technical equivalents with the help of an expressive ontology developed specifically for legal research.

One problem that the resource manager (RM) faces when enlisting the service of multiple resource agents (RAs) is that different agents may take significantly different time to return results based on the particular information sources they query. Waiting for all the results to come in before ranking and ordering all the results may not be feasible. In our current implementation, we rank and order links within each information source separately, e.g., if the user queries both the Federal statutes and Oklahoma laws we rank results returned from these two different sources separately.

The system was still slow to respond to a query, as we had to fetch the entire document, and not only the link, to find the context of the match and then rank the documents. To speed up the response, we decided to rank order only a fixed number (say N) of documents per page. When a set of N documents were fetched and ranked, they were presented to the user and the next set of documents were fetched in the background. This, unfortunately, means that rankings are consistent only among documents presented on one page, and there is no rank correlation between documents in different pages. In particular, it is not the case that documents in the first page are necessarily more relevant than documents in the second page. This might happen and is likely when the information source being queries has a good ranking mechanism to order response to queries.

The speed of operation is a limitation of the current system as we are not caching any data locally but access Internet repositories for each user request. The response time varies between a few seconds to a minute. While this may be adequate for a proof-of-concept research prototype, the system need to be redesigned and optimized for wide-scale usage.

Another shortcoming of this research is the static, hand-crafted nature of the ontology, and in particular, the frequently related words part of the ontology which is used to augment the user query. We believe that frequent word associations can be additionally gleaned from the usage of the system. This will allow the system to respond to usage patterns and modify its response to better suit the needs of the user community. A further concern is the frequency of occurrence of match based ranking scheme. While this is a reasonable heuristic, often the quality of match is more likely a semantic and not syntactic property of the document. A collaborative filtering based approach that ranks documents based on selection frequency of other users may be a more appropriate mechanism. A combination of these two metrics is probably a more practical solution with the weightage on frequency of occurrence reduced over time as more and more user interaction data becomes available.

The legal ontology developed specifically for LawBOT and based on one of the coauthors domain expertise is what makes LawBOT especially useful for non-expert users. It is a part of the LawBOT database, and is designed as a plug-and-play module. As such, from the design point of view, it is relatively straightforward to reimplement LawBOT for a different domain, say a MedicalBOT. The major work involved will be in designing the medical ontology and storing it in our database syntax. The other, more tedious, work will involve writing the Resource Agents to query medical information repositories on the internet. While this implementation may take a little time, the overall architecture and design of the system can be readily ported to a new domain.

The following are the key research and implementation issues that we foresee as our future work:

- Improving the response time of the system.
- Enhancing the ontology based search by learning further word associations based on on-line usage.
- Presenting a navigation map of web resources so that the user may restrict querying arbitrary subsets of information sources. This will also speed up the search process.
- Provide a collaborative filtering mechanism by which users will be notified of statutes, and laws in their search pattern that was liked/used by other users.
- Adding the proactive response component to the current system.

References

[1] Communications of the ACM, july 1994, volume 37, number 7, 1994. Special Issue on Intelligent Agents.

[2] Communications of the ACM, march 1997, volume 40, number 2, 1997. Special Issue on Recommender Systems.

[3] Communications of the ACM, march 1999, volume 42, number 3, 1999. Special Issue on Multiagent Systems on the Net and Agents in E-commerce.

[4] Selberg Erik and Etzioni Oren. The metacrawler architecture for resource aggregation on the web. *IEEE Expert*, 12(1):8–14, January/February 1997.

[5] Oren Etzioni. A scalable comparison-shopping agent for the world-wide web. In *Autonomous Agents Conference Proceedings*, 1997.

[6] Oren Etzioni and Mike Perkowitz. A softbot-based interface to the internet. In *CACM*, pages 72–76, 1994.

[7] Oren Etzioni and Mike Perkowitz. Intelligent agents on the internet: Fact fiction and forecast. In *IEEE Expert*, pages 44–49, August 1995.

[8] Findlaw, findlaw home page, findlaw team. http://FindLaw.com/.

[9] T. Finin, R. Fritzson, D. McKay, and R. McEntire. KQML as an agent communication language. In *Proceedings of the Third International Conference on Information and Knowledge Management*, New York, NY, 1994. ACM Press.

[10] N.R. Jennings, K. Sycara, and M. Wooldridge. A roadmap of agent research and development. *International Journal of Autonomous Agents and Multi-Agent Systems*, 1(1):7–38, 1998.

[11] Umbc kqml site. http://www.cs.umbc.edu/kqml/.

[12] H. Leiberman. Letizia: An agent that assists web browsing. In *Iternational Joint Conference on Artificial Intelligence Proceedings of 1995*, pages 924–929, 1995.

[13] Lexis Nexis Xchange (tm) homepage, lexis nexis. http://www.Lexis.com/xchange/.

[14] Law Office Information Service, Electronis Law Library (LOIS) home page. http://www.pita.com/.

[15] Morelaw, morelaw homepage, morelaw law library home page. http://www.morelaw.com.

[16] Sun microsystems page, java home page, sun microsystems, 1999. http://java.sun.com.

[17] Sun servlet page, java servlet home page, sun microsystem, 1999. http://java.sun.com.

[18] Mike Uschold and Austin Tate. Putting ontologies to use. *The Knowledge Engineering Review*, 13(1):1–3, March 1998.

[19] Westlaw, westlaw home page, westlaw team, 1999. http://westlaw.com/.