Title:

Using Citations to Facilitate Precise Indexing and Automatic Index Creation in Collections of Research Papers

Contact Author:

Shannon Bradshaw
Intelligent Information Laboratory
Northwestern University
1890 Maple Avenue
Evanston, IL 60201, USA
+1 847 467-1771 (voice)
+1 847 491-5228 (fax)
bradshaw@infolab.nwu.edu

Co-Author:

Kristian Hammond
Intelligent Information Laboratory
Northwestern University
1890 Maple Avenue
Evanston, IL 60201, USA
+1 847 467-1771 (voice)
+1 847 491-5228 (fax)
bradshaw@infolab.nwu.edu

Abstract:

We describe Rosetta, a digital library system for scientific literature. Rosetta makes it easy for people to find the information for which they are looking even when using short, imprecise queries. Rosetta indexes research articles based on the way they have been described when cited in other documents. The concise descriptions that occur in citations are similar to the short queries people typically form when searching; therefore, citations may make a better basis for indexing than do the words used within a research article itself. Using this indexing technique we are able to provide a user interface that presents users with an automatically generated directory of the information space surrounding a query. Our objective with this interface is to present people with the information for which they have asked as well as the information for which they may have intended to ask.

Keywords:

Search engines, information retrieval, reference directed indexing, citation indexing

Journal Title:

Knowledge-Based Systems

1. Introduction

In recent years, several digital library projects have established large repositories of scientific literature [6, 7, 10, 11, 21, 23]. These libraries provide coverage of fields such as Medicine, Engineering, and Computer Science by means of research articles published in a variety of books, journals, and conference proceedings.

People access these digital libraries using some sort of keyword-based query interface often provided through the Web. Unfortunately, subject searching is implemented using technology that is not designed to handle the kinds of queries people typically pose to the system. This technology is based on the vector-space model [15]. In the vector-space model, documents are represented by a list or vector of terms selected from among the most frequently used words in the document. Term selection algorithms in these systems are designed to chose terms that uniquely identify a document as much as possible within a given collection. The problem is that this technology was designed to facilitate the comparison of one document to another, but current digital libraries use it as a means of comparing requests for information (queries) to documents. When a search is performed, the description of the query is compared to term-vectors that represent the information content of documents in the collection. Those documents with term vectors matching the query are retrieved as documents that contain the requested information. However, the documents retrieved are often far from what the user is looking for, because the technique of matching queries to term-vectors is not adequate to handle the type of queries users typically pose when searching. Studies of the searching behavior of people who use digital libraries [10] and other information systems [9, 20] show that people rarely form queries longer than three words while term vectors are usually longer than twenty. As a result, queries and term-vectors are very different types of objects and comparing them is at best an imprecise means of retrieving the requested information. In addition, people rarely use features of the query interface such as the Boolean operator “and” or phrase delimiters such as quotation marks to indicate how they intend query words to be grouped together. Current indexing systems do nothing to resolve the ambiguity introduced by such queries; instead, they simply assume that people will form unambiguous queries, an assumption invalidated by the same studies on searching behavior [9, 10, 20].

In an effort to resolve these problems, we are developing a technique in which research articles are indexed using the words describing them when they are cited in other papers. We are coupling this technology with an interface in which a query retrieves a directory of the information space surrounding the query. This directory is generated automatically from the indices of documents that have been referenced using wording similar to that of the query. Our test-bed for this technology is Rosetta, an indexing and retrieval system for Computer Science research articles.

2. Rosetta

2.1 References to Research Papers

Citations from one research article to another occur for a variety of reasons. Some citations are made in support of ideas presented; others are made to cite examples of a particular viewpoint. Still others are made to name work with which the author takes exception. Recognizing and using information about the purpose of a citation is an interesting topic in itself and one that we plan to explore with future research. In our current work; however, we are concerned solely with the feature

of citations that ties the different uses together, that being that the text surrounding a citation (the reference) is usually a concise description of the information the cited document provides. For example, the paper *Agents That Reduce Work and Information Overload* by Pattie Maes is referenced by the following two sentences:

1. "There is also the question of trust, as discussed by Maes (1994)."[1]
2. "In addition, users may not trust intelligent agents since they often lack the ability to respond to user requests for clarification (Maes 1994)."[2]

These references indicate that one facet of Maes' paper is a discussion of the issue of trust between a user and an intelligent agent. Other references indicate that the agents described in this paper are agents that act to assist a person in day-to-day tasks:

1. "...computerized personal assistants which deal with meeting scheduling, e-mail filtering and re-ordering, flight booking, selection of books etc"[3]
2. "An agent that acts as a personal assistant is called an interface agent".[4]

To the human, these references serve as excellent descriptions of the ideas Maes discusses in her paper. This is because the terms "trust", "intelligent agent", and "personal assistant", are familiar and touch on specific topics, especially when used in combinations such as (trust and "intelligent agents").

References are extremely valuable as a foundation for indexing in an information system, because they pair concise, on-point descriptions of information with the documents that contain that information. As a result, the information system is much better equipped to deal with the brief and often incomplete way people typically describe an information need, because a few words is often enough to eliminate from consideration many irrelevant documents that would be retrieved by standard retrieval techniques based on content.

2.2 Building Indices from References

Rosetta indexes each document in the collection by the words and phrases used in references to that document. We refer to these words and phrases as terms. To index documents Rosetta parses the references to them and extracts noun phrases up to four words in length. We extract noun phrases because queries are typically composed of nouns and noun phrases. Noun phrases are extracted using a simple algorithm that employs the Moby Lexicon [22] as means of determining

---

[1] In *Edited Adaptive Hypermedia: Combining Human and Machine Intelligence to Achieve Filtered Information* by Kristina Hook, Asa Rudstrom, and Annika Waern.

[2] In *Instant TEA - Instant Traveling Expert Advice* by Tod Sedbrook.

[3] In *The Evolution of Intelligent Agent and Game Theory: Towards the Future of Intelligent Automation* by N. K. Khoo and Denise J. J. Chen.

[4] In *Intelligent Agents for Internet-Based Military Training* by Niraj Joshi and V. C. Ramesh.

parts of speech. This lexicon contains 230,000 English words and lists the parts of speech in which that word is used sorted from the most often used to form to the least often used. The noun phrases Rosetta extracts are those in which a noun is modified by pre-noun modifiers such as "digital library interfaces".

For the purpose of extracting noun phrases from references, Rosetta considers as a noun any word that can be used as a noun according to the lexicon and any word in a reference that is not found in the lexicon. Similarly Rosetta considers as an adjective any word that can be used as an adjective. For words that can be used as both a noun and an adjective Rosetta treats the word as a noun. Any sequence of up to four words that is consistent with the syntax of a noun phrase is extracted and used as an index for the document described by the reference.

After every document has been fully indexed, the system calculates weights for each index that represent its importance as a descriptor for each document it is used to describe. Index weights are calculated using the following metric, which is based on TFIDF [16], a standard term-weighting measure from the Information Retrieval research community:

$$w_{id} = \frac{n_{id}}{1 + \log N_i}$$

where w_{id} is the weight of index i as a label for document d, n_{id} is the number of times index i was used in reference to d, and N_i is the number of documents for which index i is used as a label.

3. Rosetta's User Interface

Recent studies of the searching behavior of people who use information retrieval systems indicate that people tend to search for information using remarkably simple queries. Jansen, et al. evaluated the usage of Excite, a popular Web search engine [9]. Excite provides a keyword search interface, in which users list the terms they expect to find in the documents they are seeking. Users may make a query more specific using the Boolean operators "and", "or", and "not" and by delimiting phrases using quotation marks. However, in an evaluation of the transaction logs of over fifty thousand queries, the authors found that people rarely make use of query language features. Instead, users typically enter a simple list of keywords as queries. Furthermore, on average these queries are less than three words in length. In a related study of the usage of Excite, Spink, et al. surveyed over three hundred users to collect, among other data, information on search strategies [20]. In this study, Spink et al. found that not one searcher used quotation marks to indicate that some or all of the terms in their query should be interpreted as a phrase, even though many searchers clearly intended that some search terms be interpreted as such. These two studies show that most users of Excite search for information by simply listing two or three terms that touch on the topics of interest. Since Excite is one of the most heavily used Web search engines it is likely that the behavior of users of Excite is representative of searching behavior on the Web in general.

Jones et al. [10] found that this behavior is not limited to the use of Web search engines, but is exhibited by users of other types of information systems as well. This study revealed that less than 30% of people searching the New Zealand Digital Library use Boolean operators and less than 20% pose queries longer than three words.

We can draw two conclusions from the results of these studies. One is that people for the most part do not make use of query language features when searching for information. They enter simple lists of words and expect the information

system to interpret them correctly. The other is that most users do not describe information needs in sufficient detail. For some people this is because they are inexperienced in using information retrieval systems. For others it is because they are not sure exactly how to describe the information for which they are looking. This may be due to lack of familiarity with the jargon of a particular field or due to the complexity of the topic of interest. Therefore, to effectively satisfy user requests for information the system must be able to interpret the query as the user intended it. It should also guide the user to those documents that satisfy an information need that is not specified in sufficient detail.

3.1 Searching for Information in Rosetta

People typically search for information by simply listing the terms that form the words and phrases that describe an information need. Therefore, the information system must infer the intended parsing of the query and respond accordingly. Rosetta accepts queries in natural language and responds based on the most reasonable interpretations of the list of words that compose a query.

When a query is submitted, if the user has not segmented the query using quotation marks and/or Boolean operators, Rosetta generates all possible interpretations of the query by parsing it into words and phrases (terms). For example, given the query:

intelligent agents trust

Rosetta finds the following parses:

1. "intelligent agents trust"
2. "intelligent agents" and trust
3. intelligent and "agents trust"
4. intelligent and agents and trust

To satisfy the request for information Rosetta selects the most specific parsing of the query that retrieves information. Query specificity is probably best illustrated by the example that "'intelligent agents' and trust" is a more specific query than, "intelligent and agents and trust". In the current collection the parsing "'intelligent agents' and trust' is selected because no documents are indexed under the phrase "intelligent agents trust".

Rosetta retrieves documents by breaking up the selected query parsing into its component terms (i.e. "intelligent agents", "trust") and retrieving the documents indexed under each term individually. It then weights each document by summing the weights of the document for each query term under which it was indexed. So any document indexed under both "intelligent agents" and under "trust" is assigned a weight that is the sum of the weight of "intelligent agents" as a descriptor for the document and the weight of "trust" as a descriptor for the document.

3.2 Finding Related Topics

For many queries simply finding those documents which best match the information need as stated is not enough, because the information need has not been described in enough detail. For these queries it is necessary to guide the user to a

more detailed description of the information for which he is looking. To do this Rosetta automatically generates a directory of the information space surrounding the information requested in a query.

In Rosetta, we build a mini-database around each term that represents the other terms used with that term in all references indexed by the system and how often the terms are used together. Each term database also contains any phrases of which the term is a part. For example, the term database for "agents" contains the term "intelligent agents". To construct the directory of related topics for a query, Rosetta locates the terms that have been used with one or more of the terms in the query. Rosetta first sorts the terms into bins based on the number of query terms with which they have been used. The terms within each bin are then sorted by the total number of times each has been used with any of the query terms. A minimal stoplist is run over the terms to eliminate terms such as, "system" and "approach". The terms are then combined with the query terms with which they have been used in the past to form topic descriptions. For example, if the query is "interface agents" Rosetta might construct related topics such as "'interface agents' and news" and "'interface agents' and mail". If the query is simply "agents" then several related topics in the form of phrases might be constructed. "Intelligent agents" and "mobile agents" are two examples. Related topics are then presented in the descending order imposed by the sort first into bins and then by usage.

3.3 Presenting Search Results

Rosetta presents search results to the user in a format that should be familiar to anyone who has used the Web. (See Figure 1.) The topics related to the query are presented as a directory at the left of the page. Each related topic is hyperlinked to pass itself to Rosetta as a query so that users may easily browse the topics suggested.

To the right of the directory, the documents matching the query are displayed. For each document, the title and authors are listed as well as a sample of the references to that document. The references provide a clear picture of how people write about a document and therefore, help the user to quickly decide which documents will satisfy his information need. Similarly, the labels listed in the directory help the user to find the group of documents most likely to contain the information for which he is looking. The user can switch to another group of documents by simply selecting the label describing that group (See Figure 2). The topic labels serve to suggest to the user how information is indexed in the system. If the first query is unsuccessful they may point him to alternative ways of describing the needed information to retrieve the documents of interest.

This interface is valuable to searchers because by suggesting less ambiguous queries this interface helps users find information sought even in the face of poorly formed queries. In addition, it helps users understand how their topic of interest fits into a body of research. Finally, by browsing the information space previously unknown work can be easily discovered.

4. System Evaluation

To test the effectiveness of Rosetta as a research tool we requested a sample of queries from the user logs of the Research Index [11] scientific literature search engine. Steve Lawrence was kind enough to oblige us. We chose to test on a sample of queries made to a similar information system, because we felt the queries would be less contrived than any we

asked users to form solely for evaluation purposes. From the sample we were sent we selected the first forty queries for which we could presume a motivating task and for which we knew Rosetta contained at least some information.

4.1 Experiment Design

A group of six people, including, one of the authors evaluated Rosetta's performance on each of the forty queries. We evaluated both the performance of Rosetta in terms of the documents it retrieved as well as the directory of related topics created in response to each query. We used a binary rating scale to evaluate each document and related topic listed; the ratings were made in the form of yes/no responses to the following questions:

- For documents: Is the document the kind of document I would want to see in response to the query?
- For related topics: Is the topic suggested a description of a useful category of information relating to the query?

We evaluated up to twenty documents and related topics for each query depending on the information Rosetta was able to find.

4.2 The Corpus

We tested Rosetta on a corpus of Computer Science research papers also lent to us by Steve Lawrence at the Research Index [12]. The corpus contains over 37,000 documents indexed using roughly 450,000 references. The references are composed of a window of about fifty words on either side of the citation, which we parse using several citation-style recognizers to extract only the words that directly refer to the cited document. See [1] for a description of our reference parsers.

4.3 Evaluating the Usefulness of Retrieved Documents

For document retrieval we are interested in a stronger measure of success than simply relevance, because a document can be considered relevant if it simply contains information on the topic(s) described in the query. Instead, we evaluated Rosetta with a measure of how well each document is likely to be useful in the task that motivated the query. Granted, we can neither be certain of the task prompting each of the queries in our selection of user log data nor did we read any paper retrieved. However, to the best of our ability we determined the motivating task and evaluated whether or not each document would be useful in that task by reading the abstracts of each document and the list of references to that document. Furthermore, we want useful documents to be retrieved among the first five presented to the user, because users typically see only the first five documents of a query response without scrolling and may evaluate the success of the query based on those first few documents. By these measures Rosetta performed very well in our evaluation. On average, 71.3% of the top five documents retrieved were determined to be useful, while 65.2% of the top 10 were determined to be useful. The very first document retrieved was useful for 76.3% of the queries.

4.4 Evaluating Information Need Satisfaction

The usefulness of retrieved documents is extremely important, but it is possible to be present a user with a large number of useful documents, but still leave that users information need partially unsatisfied. To test Rosetta's ability to completely satisfy an information need, we performed an additional, more subjective, evaluation. This evaluation is by no

means the definitive word on Rosetta's ability to satisfy information needs. For that type of result we need a user base in place, a much larger sample size, and some method of evaluating user satisfaction with Rosetta as people use it for real research tasks. In this evaluation we sought only some indication that we are on the right track with our indexing technology. We selected the first twelve of the forty queries in the test set and evaluated the total query response on a scale of 1 to 5 with each rating in the scale having the following general meaning:

1. No useful documents were retrieved.
2. Some useful documents were retrieved.
3. Several useful documents were retrieved, but the information need is largely unsatisfied.
4. Many useful documents were retrieved, but the query response is missing that one document that completely satisfies the information need.
5. The information need is completely satisfied.

Of course, it must be remembered that this evaluation was also performed using our interpretation of the information need. For the twelve queries in this evaluation the average rating was 4.13; four of the twelve query result sets were determined to completely satisfy the information need. What does this tell us about Rosetta's ability satisfy information needs? It is perhaps a small indication that we are on the right track. The real value of this evaluation was that we gained a better understanding of what composes a good result set. The four sets rated with a 5 typically contained the first paper on some technology (or one of the most important ones), documents comparing the technology to competing research, and several examples of the application of the technology. The documents retrieved for the query "RSVP" (Resource ReSerVation Protocol) are an example of set of results that seem to completely satisfy the information need described in the query. In general, these are documents that the research community has deemed important by citing them often using the words people use when searching for the information they contain. This is encouraging, because our indexing technique is based on this assumption.

4.5 Evaluating the Related Topics Suggestions

The results of our evaluation of Rosetta's ability to suggest related topics may at first glance appear to be bad; however, we are quite pleased with them. We took the simplest algorithm we could imagine, suggesting topics that are related to a query by combining query words with other words used together in references and sorting by how frequently they are used together. We use almost no knowledge about what makes a likely topic description -- a minimal stoplist only – and no knowledge about grammatical structure in references. Still, in the evaluation of our forty test queries, on average, Rosetta placed two useful related terms in the top six, and nearly two in the top five. Overall, one-third of the related topics suggested for each query were useful. Having performed this evaluation we have a much clearer picture of what a good topic suggestion should look like and equally important, how to eliminate bad suggestions. We feel that applying these lessons to the technology will yield a higher percentage of good topic suggestions.

5. Related Work

The value of using citations in documents has been explored extensively in IR research. (See [14], [19], and [24] for examples.) However, this work has been concerned mainly with what can be inferred about the similarity of documents based on an analysis of citation networks. One intuition supporting this research is that the similarity of two documents varies directly as the number of documents that cite both of them. Related to this work, is ResearchIndex (also known as CiteSeer) [12] at NEC Research. In ResearchIndex, research papers are indexed by content, but are retrieved based on the frequency with which a document has been cited. In addition, using the ResearchIndex, people can search for references matching their query as well as documents. However, to accomplish the same effect as Rosetta using Research Index is somewhat tedious.

The use of referential text as a basis for indexing is less well explored. Some Web search engines such as Google [3] and HotBot use popularity indexing [3] in which the text enclosed in anchor tags as well as content are used to index the page for which the anchor tag forms a hyperlink. However, the focus of popularity indexing is slightly different from our work in that the technology is designed to identify the most popular pages associated with query words rather than documents that have been described in a manner that most closely resembles the query. Regardless, popularity indexing is an extremely effective indexing technique and the fact that our indexing technology is so similar to that of Google bodes well for better indexing of scientific information.

Another related body of work is that of adding context to ambiguous queries. These systems are typically embedded in a document creation or browsing tasks and are not queried in the absence of such a task. One system is a version of Rosetta embedded in Emacs for use with the LaTeX document preparation system [2]. In this version of Rosetta, context is gathered from the user as he types a document. The system retrieves supporting documents on topics related to those discussed in a small window of text surrounding the cursor. Another system is Watson [4]. In Watson, Budzik has developed a similar tool for use Microsoft Word; however, Watson searches for related documents using existing Web search engines.

6. Conclusions

Rosetta uses previous descriptions of research papers found in citations to index documents so that future requests for the information those documents contain can be easily recognized and satisfied. Using this technique Rosetta is able to provide users with the documents they need in response to simple queries. Furthermore, our experiments with using references to construct lists of topics related to a query suggest that we can construct a browseable directory of the information space surrounding a query to help users who are able to recognize the information they need more easily than they are able to describe it. Our experiments, indicate that the technology implemented in Rosetta will prove valuable for research-oriented information gathering.

7. References

[1] Bradshaw, S. and Hammond, K. Constructing Indices from Citations in Collections of Research Papers. *Proceedings of the Sixty-Second Annual Meeting of the American Society for Information Science*, Washington, D.C., October 31 - November 4, 1999.

[2] Bradshaw, S. Reference Directed Indexing: Attention to the Description People Use for Information. Masters Thesis. The University of Chicago. 1998.

[3] Brin, S. and Page, L. The Anatomy of a Large-Scale Hypertextual Web Search Engine. *Proceedings of WWW '98* (Brisbane Australia, April 1998).

[4] Budzik, J., and Hammond, K. Watson: Anticipating and Contextualizing Information Needs. *Proceedings of the Sixty-second Annual Meeting of the American Society for Information Science.* Learned Information, November, 1999.

[5] Budzik, J., and Hammond, K. User Interactions with Everyday Applications as Context for Just-in-time Information Access. *Proceedings of Intelligent User Interfaces 2000*. ACM Press, 2000.

[6] Chen, H., Schatz, B., Ng, T., Martinez, J., Kirchhoff, A., and Lin, C. A parallel Computing Approach to Creating Engineering Concept Spaces for Semantic Retrieval: The Illinois Digital library Initiative Project. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Special Section on Digital Libraries: Representation and Retrieval 18, 8 (August 1996). 771-782.

[7] Davis, J. R. Creating a Networked Computer Science Technical Report Library. *D-lib Magazine*, September, 1995.

[8] Franklin, D., Bradshaw, S., and Hammond, K. (2000). Jabberwocky: You don't have to be a rocket scientist to change slides for a hydrogen combustion lecture. *Proceedings of Intelligent User Interfaces 2000.* New Orleans, LA, January 9-12, 2000.

[9] Jansen, B.J., Spink, A., Bateman, J., Saracevic, Tefko. Searchers, the Subjects they Search, and Sufficiency: A Study of a Large Sample of Excite Searches. *Proceedings of Webnet '98* (Orlando FL, November 1998). 472-477.

[10] Jones, S., Cunningham, S. J., and McNab, R. An Analysis of Usage of a Digital Library. *Proceedings of ECDL '98* (Heraklion Crete Greece, September 1998).

[11] Lawrence, S., Giles, C. L., Bollacker, K. Digital Libraries and Autonomous Citation Indexing. *IEEE Computer* 32, 6. 67-71. 1999.

[12] Lawrence, S., Bollacker, K., Giles, C. L. Indexing and Retrieval of Scientific Literature. Proceedings of the Eighth International Conference on Information and Knowledge. (Kansas City, Missouri, November 1999). 139-146.

[13] Mauldin, M. and Leavitt, J. R. R. Web Agent Related Research at the Center for Machine Translation. *Proceedings of the ACM Special Interest Group on Networked Information Discovery and Retrieval*. 1994.

[14] Marshakova, I. V. System of Document Connections Based on References (in Russian). *Nauchno-Tekhnicheskaya Informatsiya, ser*. 2, 6. 3-8. 1973.

[15] Salton, G., Wong, A., and Yang, C. S. A Vector Space Model for Automatic Indexing. *Communications of the ACM* 18, 11. 613-620. 1971.

[16] Salton, G, and Buckley, C. Term-Weighting Approaches in Automatic Text Retrieval. *Information Processing and Management*. 24, 5. 513-523. 1988.

[17] Schatz, B. R., Johnson, E. H., and Cochrane, P.A. Interactive Term Suggestion for Users of Digital Libraries: Using Subject Thesauri and Co-occurrence Lists for Information Retrieval. *Proceedings of Digital Libraries '96* (Bethesda MD, March 1996). ACM Press. 126-133.

[18] Schick, A., Gordon, L., and Haka, S. Information Overload: A Temporal Approach. *Accounting Organizations and Society*, 15 (3) 1990: pp. 199-220.

[19] Small, H. Co-citation in the Scientific Literature: A New Measure of the Relationship Between Two Documents. *Journal of the American Society for Information Science*. 24. 265-269. 1973.

[20] Spink, A., Jansen, B. J., Bateman, J. Users' Searching Behavior on the Excite Web Search Engine. *Proceedings of Webnet '98* (Orlando FL, November 1998). 828-833.

[21] The U.S. National Library of Medicine. http://www.nlm.nih.gov/nlmhome.html.

[22] Ward, Grady. A set of lexical resources. http://www.dcs.shef.ac.uk/research/ilash/Moby/

[23] Witten, I. H., and McNab, R. The New Zealand Digital Library: Collections and Experience. *The Electronic Library* 15, 6. 495-503.

[24] Yaru, D. Structural Modeling of Network Systems in Citation Analysis. *Journal of the American Society for Information Science*. 48, 10. 946-952. 1997.

Figure 1, Bradshaw, Knowledge Based Systems

Caption: Figure1: Rosetta results for the query: “agents”.

Position: Near the reference to Figure1 in section 3.3

THIS IMAGE MUST BE LARGE TO BE READABLE

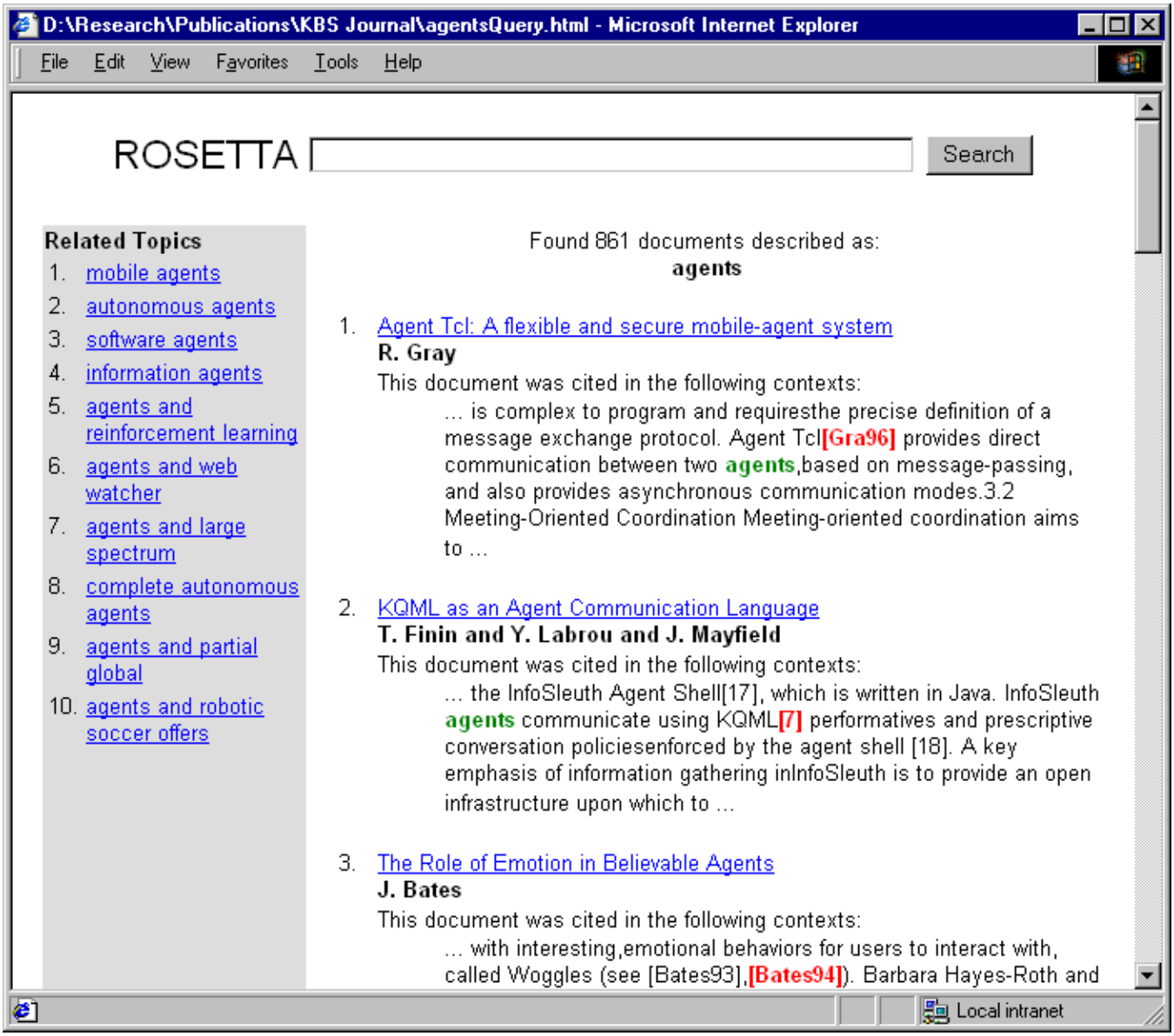

Figure 2, Bradshaw, Knowledge Based Systems

Caption: Figure 2: Rosetta after selecting the related topic: “information agents”.

Position: Near the reference to Figure2 in section 3.3

THIS IMAGE MUST BE LARGE TO BE READABLE