

# Intelligent Web Search via Personalizable Meta-Search Agents

Larry Kerschberg<sup>1</sup>, Wooju Kim<sup>2</sup>, and Anthony Scime<sup>3</sup>

<sup>1</sup>E-Center for E-Business, George Mason University

<sup>2</sup>Chonbuk National University

<sup>3</sup>SUNY-Brockport

## Abstract

This paper addresses several problems associated with the specification of Web searches, and the retrieval, filtering, and rating of Web pages in order to improve the relevance, precision and quality of search results. A methodology and architecture for an agent-based system, WebSifter is presented, that captures the semantics of a user's search intent, transforms the semantic query into target queries for existing search engines, and ranks resulting page hits according to a user-specified, weighted-rating scheme. Users create personalized search taxonomies, in the form of a Weighted Semantic-Taxonomy Tree. Consultation with a Web-based ontology agent refines the terms in the tree with positively- and negatively-related terms. The concepts represented in the tree are then transformed into queries processed by existing search engines. Each returned page is rated according to user-specified preferences such as semantic relevance, syntactic relevance, categorical match, and page popularity. Experimental results indicate that WebSifter improves the precision of web searches, thereby leading to better information.

## 1 Introduction

The World Wide Web [1] has become an essential tool to search for information. Yet, so much information is now available that people must sift and winnow through a plethora of accessible information in order to obtain meaningful information.

Typically, users initiate a keyword-based Web search by using a search engine to find documents that refer to the desired subject. Unfortunately, because of the limited ability of Web search engines to capture and interpret the user's *information needs*, many of the retrieved results may be irrelevant. There is a *semantic gap* between the user's perception of the search domain and the results provided by search engines.

To improve upon the search results obtained from Web queries, we have developed a community of search agents that result in a user-determined and experience-driven ranking of Web pages.

This is motivated by several shortcomings in available search engines:

- 1) There is a *semantic gap* between the user's perceived problem, and the keyword-based search engines;
- 2) Ranking algorithms are considered *proprietary*, so a user cannot personalize the ranking mechanism; and
- 3) Users cannot provide feedback regarding the relevance of returned pages to allow the search agent to learn user preferences.

The approach in this paper addresses these shortcomings by allowing users to specify queries as taxonomies of concepts to a meta-search agent which then enhances the terms with synonyms and antonyms, transforms the queries into formats accepted by search engines, and then performs post-

processing of the returned hits along several *relevancy components*. In addition, users may indicate whether the ranked pages are relevant to their decision-making situation.

## 2 Related Work

Most current Internet search engines suffer from *Recall* and *Precision* problems. The relatively low coverage of individual search engines led to the concept of meta-search engines, such as MetaCrawler [2], SavvySearch [3], NECI Metasearch Engine [4], and Copernic (<http://www.Copernic.com>), so as to improve the recall of a query [2]. Although coverage may improve by using a meta-search engine, the precision problem remains. Increased coverage does not necessarily imply increased precision.

Research on precision may be categorized into three major themes: content-based, collaborative, and domain-knowledge. The content-based approaches first represent a user's explicit preferences and then evaluate Web page relevance in terms of its content and user preferences. Some research takes into account Web page content and its structure to evaluate relevance [5, 6].

The collaborative approach determines information relevancy based on similarity among users rather than the similarity of information itself. There are hybrids that incorporate both above approaches. Example systems are Firefly and Ringo [7], and Siteseer [8]. The third category is the domain knowledge approach that uses domain knowledge to improve the relevancy of search results. Yahoo! uses domain knowledge to classify pages and provides a pre-defined taxonomy path. The automatic classification of Web pages into a pre-defined or a dynamically created taxonomy [9] is a related issue. Domain knowledge may be represented as a set of user-provided example Web pages [10] or as a taxonomy [11].

Another related research category is the ontology-based approach, where domain-specific ontologies are being developed for knowledge-based search. OntoSeek [12], On2Broker [13], GETESS [14], and WebKB [15] are examples of such systems. The Semantic Web [16] may eventually provide a semantically rich ontology and associated meta-tagging tools [17], enabling more powerful indexing, searching, and services [18]. At present, however, there is no common agreement on the representation of the ontology, nor the query language or reasoning mechanisms. Even so, the precision problem remains due to the huge amount of the information on the web [19].

## 3 The Web Search Decision-Making Process

WebSifter II incorporates a user-centric information relevancy evaluation scheme, which complements the above approaches. It permits the user to create a taxonomy representing his individual search intention. This taxonomy provides a context for the Web search. The taxonomy is populated with Web pages found by searches conducted using multiple search engines.

Web page rating can be viewed as a decision-making problem, where a decision maker (a user) must evaluate various alternatives (Web pages) on selected criteria (evaluation components) for his/her problem (user's Web search intention). Web page evaluation and ranking is completed using decision analytic techniques on five user-weighted evaluation components that represent different evaluation criteria.

### 3.1 Defining the User's Search Intention

The user places his information need in a context, by specifying a Weighted Semantic Taxonomy Tree (WSTT). The WSTT consists of a set of nodes that represent a concept within the user's search intention. Each node is weighted (0 to 10) to represent the importance of this concept.

Assume that a person has started a new business and needs office equipment. He wants to search for information on the web. Suppose he wants information about chairs, so he might build a query using a single term, "chair". A more skilled user of search engines, might build a query using two terms, "office" and "chair" to obtain more precise results. In this case, the term "office" provides added context for the search. The top right pane of Figure 2 shows an example of the businessman's search intention as a WSTT for the 'Office Problem'.

One drawback to using terms for keywords is that the terms may have multiple meanings. This is one of the major reasons that search engines return irrelevant search results. To address this limitation, WebSifter II allows the user to expand and refine the context using WordNet [20]. For example, chair may have any one of four meanings, each with different synonyms:

1. {chair, seat} – a seat for one person, with a support for the back,
2. {professorship, chair} – the position of professor, or a chaired professorship,
3. {president, chairman, chairwoman, chair, chairperson} – the officer who presides at the meetings of an organization,
4. {electric chair, chair, death chair, hot seat} – an instrument of death by electrocution that resembles a chair.

The user chooses meanings to represent each WSTT concept. It is assumed that the remaining concepts are not of interest, thereby obtaining both positive and negative indicators (*Positive Concept Terms* and *Negative Concept Terms*) of the user's intent. In the example, the positive concept terms are chair and seat while the others constitute the negative terms.

The WSTT schema, is translated into Boolean queries expanded by the positive concept terms. The leaf nodes of the tree denote the terms of interest to the user, and the antecedent nodes for each node form the search context. The entire tree is transformed into a set of separate queries. For the leaf path {Office, Furniture, Chair} six queries are constructed:

- (1) "Office" AND "Furniture" AND "Chair",
- (2) "Office" AND "Furniture" AND "Seat",
- (3) "Office" AND "Piece of Furniture" AND "Chair",
- (4) "Office" AND "Piece of Furniture" AND "Seat",
- (5) "Office" AND "Article of Furniture" AND "Chair",
- (6) "Office" AND "Article of Furniture" AND "Seat".

Using the AND operator provides more precision in the results. The number of queries generated from term combinations provides for coverage of possible results. Query results are stored for further processing.

## 3.2 Web Information Rating and Ranking Mechanism

WebSifter II's ranking of Web search hits by users involves the evaluation of multiple attributes, which reflect user preferences and their conception of the information need. Ranking is approached as a multi-attribute decision problem. Search results provided by multiple search engines are ranked according to decision criteria using Multi-attribute Utility Technology (MAUT) [21], Repertory Grid [22], Dimensional Analysis [23], and Analytic Hierarchy Process (AHP) [24] techniques on five components. The computed relevance values of each component are combined into a single measure of relevance.

The semantic component represents relevancy of a Web page to a user's search intent with respect to its content. The semantic relevance value of a Web page to a query is computed by counting the number of times a term appears on the page with respect to the number of terms in the associated query. Negative concepts offset the semantic relevance by adjusting for irrelevant terms on a page.

The syntactic component measures the structural aspects of the page as a function of the role of that page within the structure of a Web site. This permits an evaluation of the page independently of the specific search underway. The approach takes into account the location of the document, its role, and the well formedness of its URL [25].

The categorical match component represents the similarity measure between the structure of the user-created taxonomy and search engine category information for the retrieved Web page. Many popular search engines, for example Yahoo! and Google, respond to user's queries not only with a list of URLs but also with categorical information for each Web page. Such categorical information helps users filter results. The categorical match component is designed to provide the benefits of manual filtering by automatic means.

The Search Engine component represents the user's biases toward and confidence in a search engine's results. A user preference value is assigned to each search engine.

The number of requests for the specific page measures popularity. There are several publicly available popularity services, which are accessed by the popularity component.

Finally, after results are returned the user may review the ranked pages and indicate page relevance to the WTTS. Relevancy selection drives a feed-forward neural network mechanism learning more about the search intent, dynamically re-rating and re-ranking the results list [26].

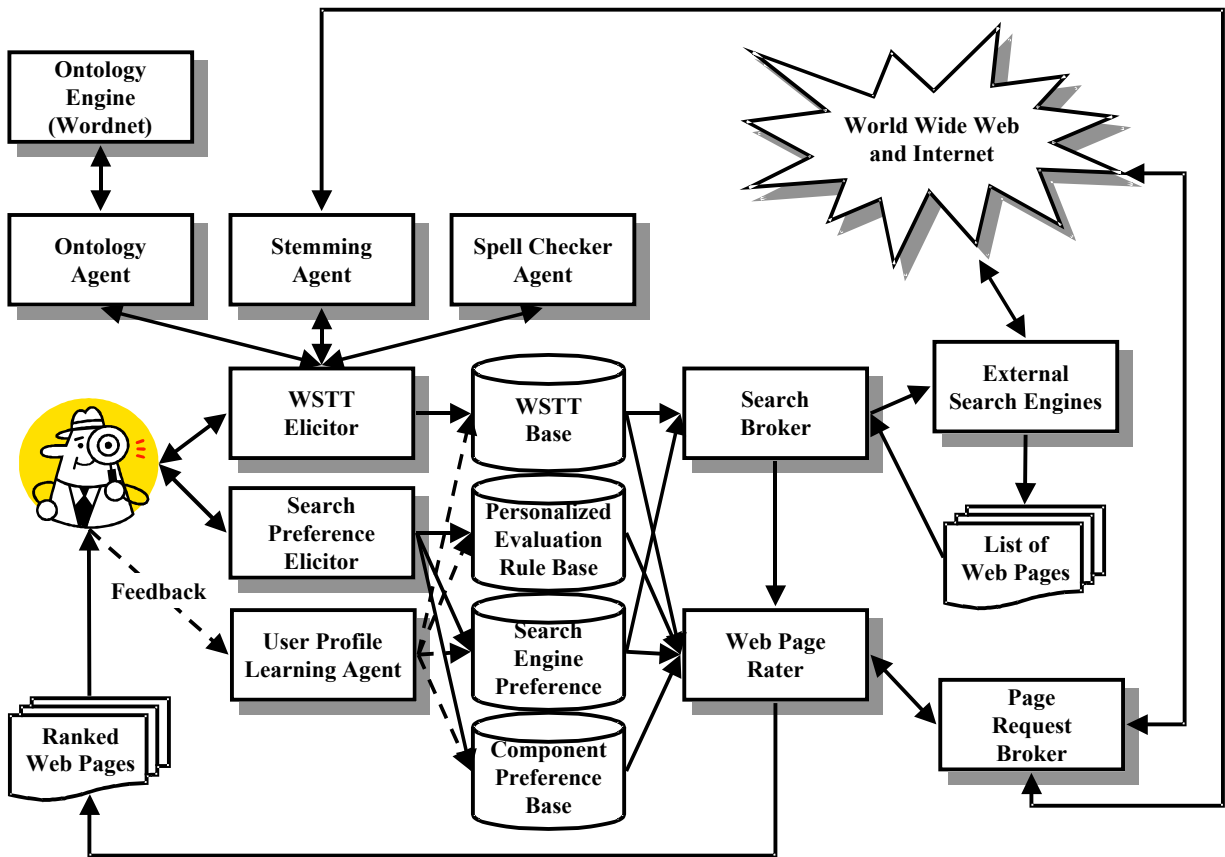
## 4. WebSifter II System Architecture

WebSifter II is a semantic taxonomy-based personalizable meta-search agent system. Figure 1 shows the overall architecture of WebSifter II. The WSTT elicitor guides the user in building the taxonomy tree, assigning weights on each node, and choosing node meanings found by the ontology agent. The WSTT is stored in XML format. The stemming agent transforms terms in a concept Web page content into stemmed terms. The search preference elicitor captures the user's search preferences. The user expresses his search preference by assigning preference weights to each of the preference components, syntactic classes, and search engines.

The search broker interprets the WSTT to generate query statements. It queries information from popular search engines and stores the results.

The page request broker obtains the content of a specific URL. The Web page rater evaluates Web pages and displays ranked results to the user. The user profile-learning agent allows the user to

provide feedback on the relevancy of the proposed Web page hits, learns about user’s search preferences, and updates the user profile.



**Figure 1: System Architecture for WebSifter II**

#### 4.1 Implementation

WebSifter II is implemented as a working Java prototype, except for the spell checker. Figure 2 shows the main screen and illustrative results for a search query. The top-left pane contains specified WSTT queries with “Office Problem” highlighted. The top-right pane shows the Weighted Semantic Taxonomy Tree construct by the user for this search.

The bottom pane shows results ranked by Total Relevance. The bottom pane left-most column is provided to obtain user feedback regarding page relevance to the search.

#### 4.2 Experimental results

Results from three experiments demonstrate the individual effects of facets of WebSifter. Results of expanding the search through positive and negative concepts are shown in Table 1 of Figure 3. The page hits retrieved by WebSifter II for the search of a just a single term, “chair” together with the

selected concept terms of seat for a person. In this table, the WebSifter ranking of web pages appears in the first column.

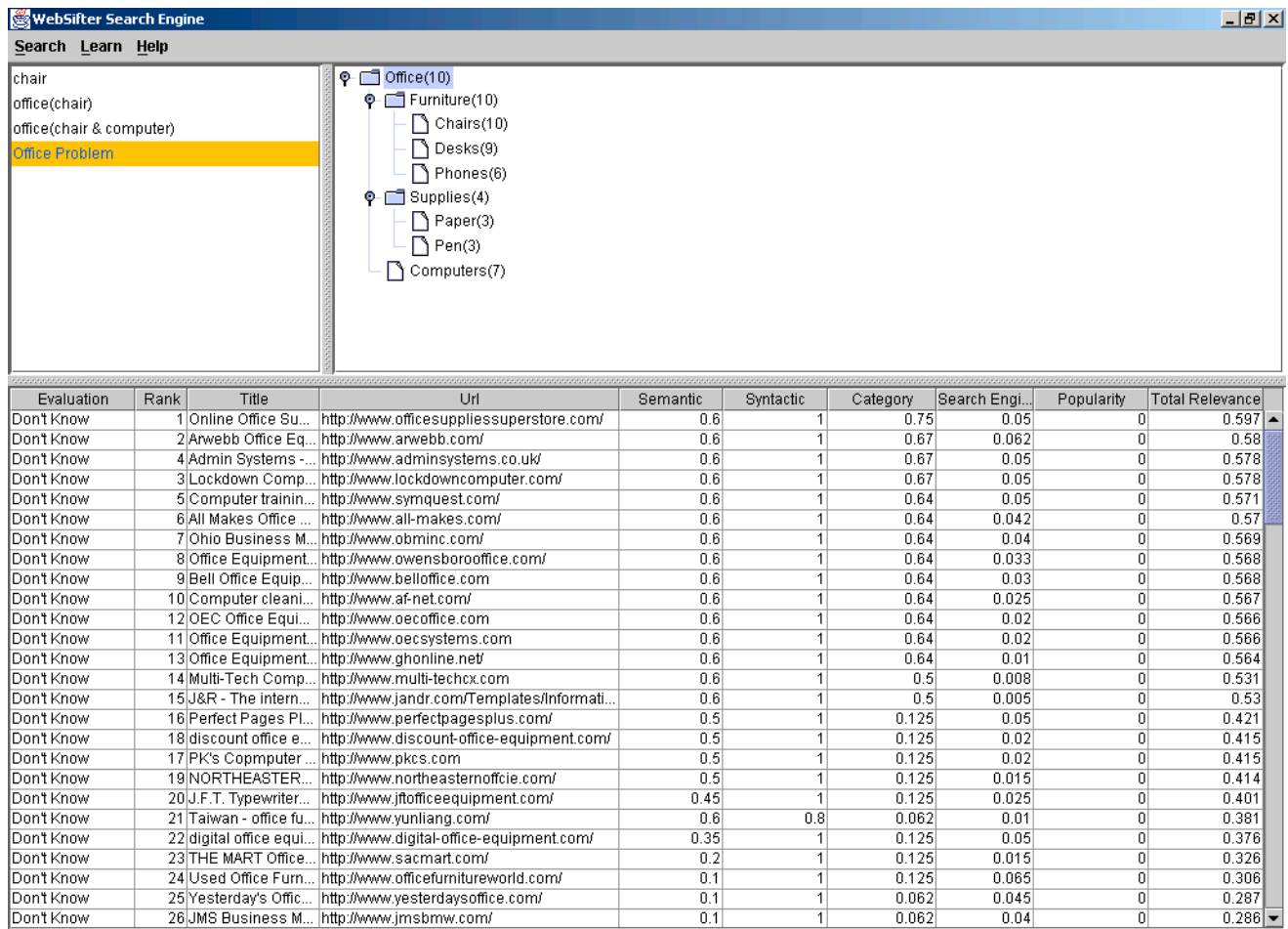


Figure 2: A Screen Shot of Web Page Rater Results

The five columns on the right compare the WebSifter ranking to those of various search engines. Note that many of WebSifter's highly-ranked pages do not appear in the top-twenty rankings of other search engines; this is due to WebSifter's use of *semantic concepts* from WordNet to enhance the search terms. The relevancy decisions are based on the user's subjective assessment as to whether the corresponding Web page is relevant to a real chair. WebSifter found six relevant Web pages.

Tables 2 and 3 of Figure 3 show the retrieved page-hits when the taxonomy is extended to "Office and Chair". In Table 2 WebSifter has a 95%-hit ratio with the only irrelevant page-hit ranked 20<sup>th</sup>. This is due to the influence of the category match component, and the 'w/o Categ' column shows the corresponding WebSifter ranking with the category match component suppressed.

Table 3 of Figure 3 shows the WebSifter retrieved page-hits and their ranks for the case where the categorical match is turned off initially. The right-most column labeled 'with Categ' shows what the WebSifter ranking would have been were the category match component activated. Note that the non-relevant page hits would have been ranked 20<sup>th</sup> and below. The results in the three tables indicate that the concept taxonomy and the categorical match component affect considerably the resulting rankings. The category match component provides a 15% performance improvement in the precision to the concept taxonomy.

**Table 1: Results from Websifter for the Query “Chair” and Comparison with Other Search Engines**

Rank	URL	Relevancy	Copernic	AltaVista	Google	Yahoo	Excite
1	www.countryseat.com	Y	-	-	-	-	-
2	www.infant-car-seat.com/	N	-	-	-	-	-
3	www.chairmaker.co.uk/	Y	-	-	-	19	-
4	www.convertible-car-seat.com/	N	-	-	-	-	-
5	www.booster-car-seats.com/	N	-	-	-	-	-
6	www.booster-seats-online.com/	N	-	-	-	-	-
7	www.booster-car-seat.com/	N	-	-	-	-	-
8	www.podiatrychair.com/	N	-	-	-	-	9
9	www.carolinachair.com/	Y	-	-	-	9	-
10	www.chairdancing.com/	N	-	-	-	-	-
11	www.massage-chairs-online	N	-	-	-	-	13
12	www.panasonic-massage-	N	-	-	-	-	14
13	www.fairfieldchair.com/	Y	-	-	15	-	-
14	www.gasserchair.com/	Y	-	16	-	-	-
15	www.chairtech.com/	Y	-	18	-	-	-
16	www.snugseat.com/	N	-	-	-	-	-
17	www.seat.com/	N	-	-	-	-	-
18	www.fifthchair.org/	N	3	2	5	8	3
19	www.painted-	N	19	-	9	-	-
20	www.jeanmonnetprogram.org/	N	5	1	1	-	5

Notes:

Y – relevant N – irrelevant

Numbers indicate rank order from search engine ‘-’ pages ranked lower than 20<sup>th</sup> or where not found by the search engine

**Table 2: Office Chair Taxonomy with Categorical Match**

Rank	URL	Relevancy	w/o Categ
1	www.seatingvfm.com/	Y	1
2	www.officechair.co.uk/	Y	2
3	www.AmericanErgonomics.com/	Y	9
4	www.ompchairs.com/	Y	22
5	www.klasse.com.au/	Y	4
6	www.cyberchair.com/	Y	46
7	www.leap-chair.com	Y	47
8	www.seizaseat.com/	Y	50
9	www.zackback.com	Y	49
10	www.fairfieldchair.com	Y	2
11	www.chair-ergonomics.com/	Y	5
12	www.buy-ergonomic-chairs.com/	Y	6
13	www.jfainc.com/	Y	7
14	www.chairtech.com/	Y	8
15	www.plasticfoldingchairs.com/	Y	13
16	www.kneelsit.com/	Y	10
17	www.home-office-furniture-store.com/	Y	11
18	www.home-office-furniture-site.com/	Y	12
19	www.amadio.it/uk/	Y	19
20	www.newtrim.co.uk/	N	15

**Table 3: Office Chair Taxonomy without Categorical Match**

Rank	URL	Relevancy	with Categ
1	www.seatingvfm.com/	Y	1
2	www.fairfieldchair.com	Y	10
3	www.officechair.co.uk/	Y	2
4	www.klasse.com.au/	Y	5
5	www.chair-ergonomics.com/	Y	11
6	www.buy-ergonomic-chairs.com/	Y	15
7	www.jfainc.com/	Y	13
8	www.chairtech.com/	Y	14
9	www.AmericanErgonomics.com/	Y	3
10	www.kneelsit.com/	Y	16
11	www.home-office-furniture-store.com/	Y	17
12	www.home-office-furniture-site.com/	Y	18
13	www.plasticfoldingchairs.com/	Y	15
14	www.office-interior-plans.com/	N	21
15	www.newtrim.co.uk/	N	20
16	www.oa-chair.com/	N	23
17	www.buy-ergonomic-chair.com/	Y	24
18	www.countryseat.com/	Y	26
19	www.amadio.it/uk/	Y	19
20	www.mobile-office-desk.com/	N	28

**Figure 3: WebSifter Experimental Results**

## 5. Conclusions

We have presented a methodology and architecture for a semantic taxonomy-based personalizable meta-search agent. WebSifter II achieves two important and complementary goals: it allows users more expressive power in formulating their Web searches, and it improves the relevancy of search results based on the user's real intent. In contrast to previous research, we have focused not only on the search problem itself, but also on the decision-making problem that motivates users to search the Web.

The Weighted Semantic Taxonomy Tree provides a mechanism for users to specify the context and intent of domain-specific terms related to the search problem. User may also state their preferences and weights for the five components by which a Web page is evaluated: semantic relevance, syntactic relevance, categorical match, search engine preference, and page popularity.

To improve the precision of the retrieved information, WebSifter uses a hybrid rating scheme that considers both the user's search intent as represented by the WSTT and the user's search preference represented by preferences associated with the relevance components.

Experimental results indicate that this approach improve the precision of the search results, and allows users to control the search terms as well as the overall ranking process. This approach could be augmented with ontologies developed by organizations to provide taxonomies for employees to use in knowledge-directed searches. This is a topic of ongoing research by the WebSifter group.

## Acknowledgements

This research was sponsored, in part, by a grant from the Virginia Center for Innovative Technology to the E-Center for E-Business, <http://eceb.gmu.edu>, at George Mason University. The authors wish to acknowledge Hanjo Jeong and Srikala Kanumuru for their Java-based implementation of the WebSifter II prototype.

## 6 References

- [1] T. Berners-Lee, R. Cailliau, A. Loutonen, H. F. Nielsen, and A. Secret, "The World-Wide Web," *Communications of the ACM*, vol. 37, pp. 76—82, 1994.
- [2] E. Selberg and O. Etzioni, "The MetaCrawler Architecture for Resource Aggregation on the Web," *IEEE Expert*, vol. 12, pp. 11-14, 1997.
- [3] A. E. Howe and D. Dreilinger, "Savvy Search: A Metasearch Engine that Learns which Search Engines to Query," *AI Magazine*, vol. 18, pp. 19-25, 1997.
- [4] S. Lawrence and C. L. Giles, "Context and Page Analysis for Improved Web Search," *IEEE Internet Computing*, vol. 2, pp. 38-46, 1998.
- [5] S. Chakrabarti, B. Dom, and P. Indyk, "Enhanced Hypertext Categorization using Hyperlinks," presented at Proceedings of ACM SIGMOD International Conference on Management of Data, Seattle, Washington, 1998.
- [6] Y. Li, "Toward a Qualitative Search Engine," *IEEE Internet Computing*, vol. 2, pp. 24-29, 1998.
- [7] P. Maes, "Agents that reduce work and information overload," *Communications of the ACM*, vol. 37, pp. 30-40, 1994.

- [8] K. D. Bollacker, S. Lawrence, and L. Giles, "Discovering Relevant Scientific Literature on the Web," *IEEE Intelligent Systems*, vol. 15, pp. 42-47, 2000.
- [9] H. Chen and S. Dumais, "Bringing Order to the Web: Automatically Categorizing Search Results," presented at Proceedings of the CHI 2000 conference on Human factors in computing systems, The Hague Netherlands, 2000.
- [10] Y. Aridor, D. Carmel, R. Lempel, A. Soffer, and Y. S. Maarek, "Knowledge Agent on the Web," presented at Proceedings of the 4th International Workshop on Cooperative Information Agents IV, 2000.
- [11] S. Chakrabarti, M. v. d. Berg, and B. Dom, "Focused Crawling: A New Approach to Topic-Specific Web Resource Discovery," presented at Proceedings of the Eighth International WWW Conference, 1999.
- [12] N. Guarino, C. Masolo, and G. Vetere, "OntoSeek: Content-based Access to the Web," *IEEE Intelligent Systems*, vol. 14, pp. 70-80, 1999.
- [13] D. Fensel, J. Angele, S. Decker, M. Erdmann, H.-P. Schnurr, S. Staab, R. Studer, and A. Witt, "On2broker: Semantic-Based Access to Information Sources at the WWW," presented at Proceedings of the World Conference on the WWW and Internet (WebNet 99), Honolulu, Hawaii, USA, 1999.
- [14] S. Staab, C. Braun, I. Bruder, A. Dusterhoft, A. Heuer, G. Neumann, B. Prager, J. Pretzel, H.-P. Schnurr, R. Studer, and H. Uszkoreit, "A System for Facilitating and Enhancing Web Search," presented at Proceedings of IWANN '99 - International Working Conference on Artificial and Natural Neural Networks, Berlin, Heidelberg, 1999.
- [15] P. Martin and P. W. Eklund, "Knowledge Retrieval and the World Wide Web," *IEEE Intelligent Systems*, vol. 15, pp. 18-25, 2000.
- [16] J. Hendler, "Agents and the Semantic Web," *IEEE Intelligent Systems*, vol. March/April 2001, pp. 30-37, 2001.
- [17] N. F. Noy, M. Sintek, S. Decker, M. Crubézy, R. W. Ferguson, and M. A. Musen, "Creating Semantic Web Contents with Protégé-2000," *IEEE Intelligent Systems*, vol. March/April 2001, pp. 60-71, 2001.
- [18] S. A. McIlraith, T. C. Son, and H. Zeng, "Semantic Web Services," *IEEE Intelligent Systems*, vol. March/April 2001, pp. 46-53, 2001.
- [19] E. J. Glover, S. Lawrence, M. D. Gordon, W. P. Birmingham, and C. L. Giles, "Web Search - Your Way," *Communications of the ACM*, vol. 44, pp. 97-102, 2001.
- [20] G. A. Miller, "WordNet a Lexical Database for English," *Communications of the ACM*, vol. 38, pp. 39-41, 1995.
- [21] D. A. Klein, *Decision-Analytic Intelligent Systems: Automated Explanation and Knowledge Acquisition*: Lawrence Erlbaum Associates, 1994.
- [22] J. H. Boose and J. M. Bradshaw, "Expertise Transfer and Complex Problems: Using AQUINAS as a Knowledge-acquisition Workbench for Knowledge-Based Systems," *Int. J. Man-Machine Studies*, vol. 26, pp. 3-28, 1987.

- [23] M. E. Martin, *Analysis and Design of Business Information Systems*. Englewood Cliffs, NJ: Prentice hall, 1991.
- [24] T. L. Saaty, *The Analytic Hierarchy Process*. New York: McGraw-Hill, 1980.
- [25] A. Scime and L. Kerschberg, "WebSifter: An Ontology-Based Personalizable Search Agent for the Web," presented at International Conference on Digital Libraries: Research and Practice, Kyoto Japan, 2000.
- [26] W. Kim, L. Kerschberg, and A. Scime, "Personalization in a Semantic Taxonomy-Based Meta-Search Agent," presented at International Conference on Electronic Commerce 2001 (ICEC 2001), Vienna, Austria, 2001.