

# WebSifter II: A Personalizable Meta-Search Agent based on Semantic Weighted Taxonomy Tree

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

1) E-Center for E-Business, George Mason University  
4400 University Drive, Fairfax, VA 22030, USA  
kersch,wkim1@gmu.edu

2) Department of Computer Science, SUNY-Brockport  
ascime@brockport.edu

## Abstract

This paper addresses the problem of specifying, retrieving, filtering and rating Web searches so as to improve the relevance and quality of hits, based on the user's search intent and preferences. We present a methodology and architecture for an agent-based system, called WebSifter II, that captures the semantics of a user's decision-oriented search intent, transforms the semantic query into target queries for existing search engines, and then ranks the resulting page hits according to a user-specified weighted-rating scheme. Users create personalized search taxonomies via our Weighted Semantic-Taxonomy Tree. The terms in the tree can be refined by consulting a web taxonomy agent such as Wordnet. The concepts represented in the tree are then transformed into a collection of queries processed by existing search engines. Each returned page is rated according to user-specified preferences such as semantic relevance, syntactic relevance, categorical match, page popularity and authority/hub rating.

## 1. Introduction

With the advent of Internet and WWW, the amount of information available from the web grows exponentially every day. However, having too much information at one's fingertips does not always mean good quality information, and rather, it may often prevent a decision maker from making sound decisions, usually degrading the quality of decision.

Although search engines assist users in finding information, many of the results are irrelevant to the decision problem. This is due in part, to the keyword search approach, which does not capture the user's intent. Search engines also have their own ranking system, which a user's criteria may change over time as more information about the problem is gathered. Thus, there is a "semantic gap" between the user's perception of the problem domain and the search results provided by search engines.

To overcome these two major problems, we proposed a weighted semantic taxonomy-based personalizable meta-search agent approach. We build upon the ideas presented by Scime and Kerschberg [1]. We develop a tree-structured search intent representation scheme with which users describe their search intent. We also present an elaborate user preference representation scheme based on various components, each of which represents a specific decision-criterion. In order to rate the relevance of a page hit, we use a decision-analytic rating mechanism combining the WSTT and the component-based preference representation. Finally, we have designed and are presently implementing a meta-search agent called WebSifter II that cooperates with Wordnet for concept retrieval, and most well-known search engines.

## 2. Related Work

Most of current internet search engines such as Yahoo, Excite, Altavista, WebCrawler, Lycos, Google, etc. suffer from *Recall* and *Precision* problems [2]. The relatively low coverage of individual search engines leads to the use the concept of meta-search engines to improve the recall of a query. Examples are MetaCrawler [3], SavvySearch [4], NECI Metasearch Engine [5], and Copernic (<http://www.copernic.com>). This meta-search engine approach partly addresses the recall problem but still suffers from the precision problem.

We can categorize research regarding the precision problem into three major themes: content-based, collaborative, and domain-knowledge approaches.

The content-based approach first represents a user's explicit preferences and then evaluates web page relevance in terms of its content and user preferences. Syskill & Webert [6], WebWatcher [7], WAWA [8], and WebSail [9] fall into this category.

The collaborative approach determines information relevancy based on similarity among users rather than similarity of information itself. Example systems are Firefly and Ringo [10], Phoaks [11], and SiteSeer [12]. In addition, some hybrid approaches incorporate both approaches for example Fab [13], Lifestyle Finder [14], WebCobra [15].

The third category is the domain knowledge approach that uses user and organizational domain knowledge to improve the relevancy of search results. One of the popular domain knowledge approaches provides a predefined taxonomy path, e.g., Yahoo!. Some research incorporates users domain knowledge in a more explicit way [16, 17].

From this survey of related research, we have identified several aspects that merit further consideration. First, most approaches force users to use a search engine in a passive rather than active manner. Second, current approaches lack sufficient expressive power to capture a users' search intent and preferences. Third, most approaches do not take full advantage of domain-specific knowledge with which to scope the search, interpret, and classify the query result.

Regarding the first limitation, there is another related research category, the ontology-based approach such as OntoSeek [18], On2Broker [19], and WebKB [20].

Although the ontology-based approach is a promising way to solve some aspects of the precision problem, it still requires major pre-requisites and even if these prerequisites are satisfied, the precision problem in web search will remain due to the huge amount of the information on the web. That is, a user-centric information relevancy evaluation scheme will complement the above approaches.

### 3. Weighted Semantic Taxonomy-Tree-Based Approach

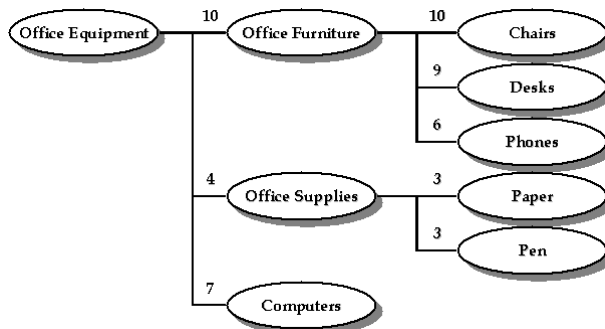
#### 3.1 Weighted Semantic Taxonomy Tree

Usually a keyword-based search representation is insufficient to express a user's search intent. By postulating a user's decision-making process, we can support readily query formulation and search. In our approach, we represent a user's search intent by a hierarchical concept tree with weights associated with each concept, thereby reflecting user-perceived relevance of concepts to the search. We call this the *Weighted Semantic Taxonomy Tree* (WSTT) model.

Figure 1 shows a realistic example of the businessman's search intention using our WSTT scheme. We can translate the upper sub-tree of Figure 2 as that a businessman wants to find information about chairs, desks, and phones within the context of office furniture and office equipment where the numbers that appear to the left of each term, 10, 9, and 6 denote the respective importance levels of chairs, desks, and phones.

One drawback of using simple terms is that it may have multiple meanings, which is one of the major reasons that search engines return irrelevant search results. To address this limitation, we introduce the notion of "word senses" from Wordnet [21] into our WSTT scheme to allow users to refine their search intention and

the user can choose one of the concepts available from Wordnet for the term of a specific node in WSTT.



**Figure 1 An Example of WSTT that represent a businessman's search intention**

For example, the "chair" term has the following four possible concepts from Wordnet.

- (1) {chair} // a seat for one person,
- (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, and
- (4) {electric chair, chair, death chair, hot seat} // an instrument of death by electrocution that resembles a chair.

If the user wants to search for a chair to sit on, he would choose the first concept. If the user selects the first concept, then without loss of generality, we can assume that the remaining concepts are not of interest, thereby obtaining both positive and negative indicators of his intent. Now, let's distinguish the set of terms of selected concept from the set of terms of the unselected concepts as *Positive Concept Terms* and *Negative Concept Terms*. We take both concept terms into account when we compute the relevancy of a web page with respect to a WSTT to achieve a better relevancy.

#### 3.2 Multi-Attribute-Based Search Preference Representation

The ranking of web search hits by users involves the evaluation of multiple attributes, which reflect user preferences and their conception of the decision problem. In our approach, we pose the ranking problem as a multi-attribute decision problem. Thus, we examine the search results provided by multiple search engines, and rank the pages, according to multiple decision criteria. Both Multi-Attribute Utility Technology (MAUT) [22] and Repertory Grid [23] are two major approaches that address our information evaluation problem. Our ranking approach combines MAUT and the Repertory Grid. We define six search evaluation components as follows:

- (1) *Semantic* component: represents a web page's relevance with respect to its content.
- (2) *Syntactic* component: represents the syntactic

relevance with respect to its URL.

- (3) *Categorical Match* component: represents the similarity measure between the structure of user-created taxonomy and the category information provided by search engines for the retrieved web pages.
- (4) *Search Engine* component: represents the user's biases toward and confidence in search engine's results.
- (5) *Authority/Hub* component: represents the level of user preference for *Authority* or *Hub* sites and pages [24].
- (6) *Popularity* component: represents the user's preference for popular sites.

Further, in this multi-component-based preference representation scheme, the user can assign a preference level to each of these components, and also to each available search engine within the search engine component. Figure 2 conceptually depicts our scheme, where each number assigned to an edge denotes user's preference level for that component. This multi-component preference scheme allows users more control over their searches and the determination of a page's relevance.

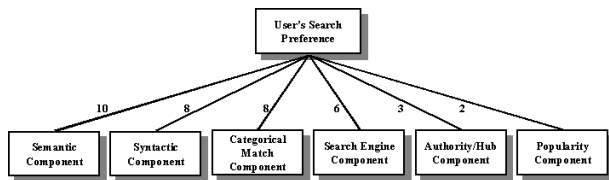


Figure 2 A Conceptual Model of User's Preference Representation Scheme

### 3.3 Gathering Web Information based on Search Intention

At present, there is no search engine that accepts a search request based on the WSTT. We have developed a translation mechanism from our WSTT-based query, to Boolean queries that most of current search engines can process. First, we transform the entire WSTT tree like Figure 1 into a set of separate queries where each is acceptable to existing search engines. To do this, first we decompose the tree into a set of paths from the root to each leaf node. Then for each path, we generate all possible combinations of terms, when selecting one term from the positive concept terms of each node in the path from a root node to a leaf node. Finally, we pose each query to search engines to obtain query results and the query results are stored for further processing, as discussed in the next section.

### 3.4 Unified Web Information Rating Mechanism

In this section, we briefly discuss a rating mechanism to evaluate each resulting page hit from the target search engines for the generated query statements. Through this mechanism, each web page will have its own value representing the relevance level from the user's

viewpoint. To accomplish this goal, six relevance values of a web page are computed, corresponding to each of the six components. Then a composite value of these six relevance values is computed based on a function of the multi-attribute-based search preference representation scheme. We leave out the detailed discussion of how to compute this composite relevance value and a set of methods to compute each of component's relevance values for want of space and the readers can refer to [25].

## 4. System Architecture of WebSifter II

In this section we present the architecture of WebSifter II, a semantic taxonomy-based personalizable meta-search agent system. Figure 3 shows the overall architecture of WebSifter II and its components. Major information flows are also depicted. WebSifter II consists of eight subsystems and four major information stores.

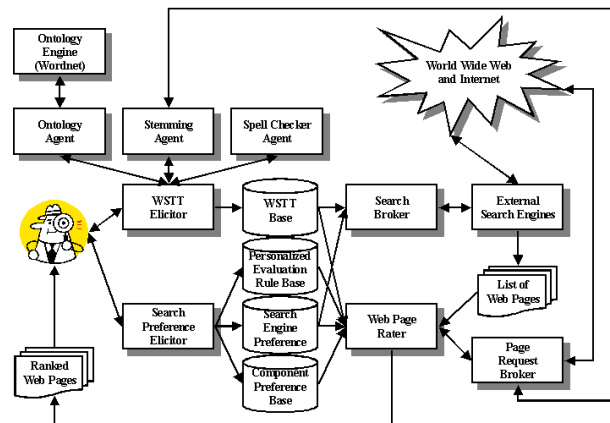


Figure 3 System Architecture of WebSifter II

Now let's briefly introduce each of the components, their roles, and related architectural issues.

#### 1) WSTT Elicitor

The WSTT elicitor supports the entire processes required in the section 3.1 to build a WSTT in a GUI environment. A user can express his search intent as a WSTT through interactions with the WSTT elicitor. This includes building a taxonomy tree, assigning weights on each node, and choosing a concept from available list of Wordnet concepts. To achieve this goal, the WSTT elicitor also cooperates with ontology agent, stemming agent, and spell check agent. Once a user finishes building a WSTT, then WSTT elicitor stores the WSTT information into the WSTT base in XML format.

#### 2) Ontology Agent

The ontology agent is responsible for requesting available concepts of a given term via a web version of Wordnet and also interpreting the corresponding HTTP based results. The agent receives requests for the concepts from WSTT elicitor and returns available concepts in an understandable form.

#### 3) Stemming Agent

Our stemming agent is developed based on Porter's algorithm [26]. It has two major roles: 1) to cooperate with WSTT elicitor in transforming the terms in a concept to the stemmed terms, and 2) to transform the content of web pages into the stemmed terms internally through cooperation with a page request broker. As a result, the terms in concepts and the terms in web pages can be compared to each other via their stemmed versions.

#### 4) Spell Check Agent

Spell check agent monitors user's text input to the WSTT elicitor and checks and suggests correct words to the user in real time.

#### 5) Search Preference Elicitor

Search preference elicitor, via a GUI, supports the process required in section 3.2 to capture the user's search preferences. A user can express his search preference through interaction with this search preference elicitor by assigning their preference weights to each of preference components and also to their favorite search engines. Moreover, it allows the user to modify the default values assigned to the web page classification rules used in computing the relevancy of syntactic component. Whenever the user modifies them, it instantly updates the related information stored in the Personalized Evaluation Rule Base, the Search Engine Preference Base, and the Component Preference Base.

#### 6) Search Broker

Search broker performs the processes required in section 3.3. It first interprets the XML-based WSTT and then generates all corresponding query statements. Using this set of queries, it requests information from a set of popular search engines simultaneously. Finally, it interprets the results returned from the search engines and then stores parsed information in a temporary data store.

#### 7) Page Request Broker

Page request broker is responsible for requesting the content of a specific URL and it cooperates with both the stemming agent and the web page rater.

#### 8) Web Page Rater

Web page rater supports the entire web page evaluation process required in section 3.4 and also is responsible for displaying the evaluation result to the users. This subsystem is the most complex and computationally intensive module of WebSifter II, and it uses all of four major information stores and also communicates with search broker and page request broker.

### 5. System Implementation

We are currently developing our meta-search agent system, WebSifter II. Some of sub-systems such as the ontology agent, stemming agent, search broker and page request broker are already developed and are operational. We have almost finished the development of the WSTT elicitor, while the search preference elicitor and the web page rater are still under development. We also plan to

incorporate a commercial spell check agent into our system.

Figure 4 shows an illustrative screen where the user is building WSTT using WSTT elicitor. Figure 5 shows another screen of the WSTT elicitor supporting the selection of an intended concept from available concepts for a given term that have been obtained through cooperation with the ontology agent.

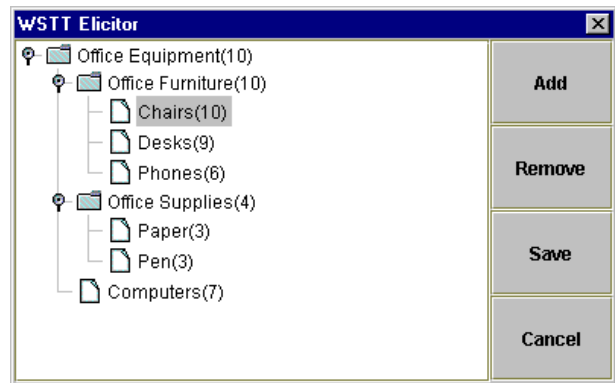


Figure 4 An Illustrative Screen of WSTT Elicitor

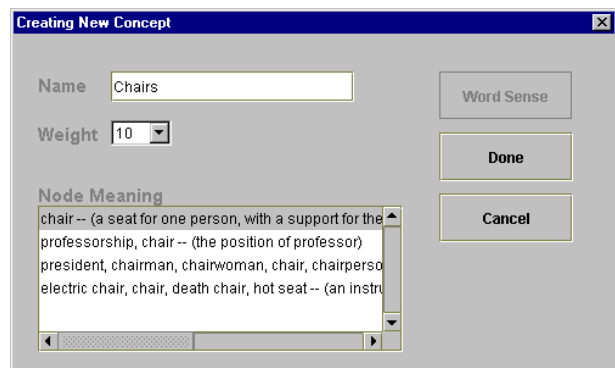


Figure 5 An Illustrative Screen for Concept Creation

### 6. Conclusions

We have proposed a semantic taxonomy-based personalizable meta-search agent approach to achieve two important and complementary goals: 1) allowing users more expressive power in formulating their web searches, and 2) improving the relevancy of search results based on the user's real intent. Now, let's briefly summarize what we have done with three concluding remarks as follows.

First, to enhance user's search intent and preference expressional power, we propose a search-intention representation scheme, the Weighted Semantic-Taxonomy Tree, by which users express their real search intentions by specifying domain-specific concepts, assigning appropriate weights to each concept, and expressing their decision problem as a structured tree of concepts. We also allow users to express their search result evaluation preferences as a function of six preference components.

Second, to enhance the *precision* of the retrieved information, we present a hybrid rating mechanism which considers both the user's search intent represented by the WSTT and user's search preference represented by multi-preference components.

Third, we have designed and are presently implementing a meta-search agent system called WebSifter II that cooperates with Wordnet for concept retrieval, and most well known search engines for web page retrieval. For the empirical validation of our approach, we are also doing some real world experiments of our system.

### References

- [1] Scime, A. and L. Kerschberg, "WebSifter: An Ontology-Based Personalizable Search Agent for the Web," *International Conference on Digital Libraries: Research and Practice*, Kyoto Japan, 2000, pp. 493-446.
- [2] Lawrence, S. and C. L. Giles, "Accessibility of Information on the Web," *Nature*, vol. 400, 1999, pp. 107-109.
- [3] Selberg, E. and O. Etzioni, "The MetaCrawler architecture for resource aggregation on the Web," *IEEE Expert*, vol. 12, no. 1, 1997, pp. 11-14.
- [4] Howe, A. E. and D. Dreilinger, "Savvy Search: A Metasearch Engine That Learns Which Search Engines to Query," *AI Magazine*, vol. 18, no. 2, 1997, pp. 19-25.
- [5] Lawrence, S. and C. L. Giles, "Context and page analysis for improved Web search," *IEEE Internet Computing*, vol. 2, no. 4, 1998, pp. 38-46.
- [6] Ackerman, M., et al., "Learning Probabilistic User Profiles - Applications for Finding Interesting Web Sites, Notifying Users of Relevant Changes to Web Pages, and Locating Grant Opportunities," *AI Magazine*, vol. 18, no. 2, 1997, pp. 47-56.
- [7] Armstrong, R., et al., "WebWatcher: A Learning Apprentice for the World Wide Web," *Proceedings of the 1995 AAAI Spring Symposium on Information Gathering from Heterogeneous, Distributed Environments*, 1995.
- [8] Shavlik, J. and T. Eliassi-Rad, "Building intelligent agents for web-based tasks: A theory-Refinement approach," *Proceedings of the Conference on Automated Learning and Discovery: Workshop on Learning from Text and the Web*, Pittsburgh, PA, 1998.
- [9] Chen, Z., et al., "WebSail: from on-line learning to Web search," *Proceedings of the First International Conference on Web Information Systems Engineering*, vol. 1, 2000, pp. 206-213.
- [10] Maes, P., "Agents that reduce work and information overload," *Communications of the ACM*, vol. 37, no. 7, 1994, pp. 30-40.
- [11] Terveen, L., et al., "PHOAKS: a system for sharing recommendations," *Communications of the ACM*, vol. 40, no. 3, 1997, pp. 59-62.
- [12] Bollacker, K. D., et al., "Discovering Relevant Scientific Literature on the Web," *IEEE Intelligent Systems*, vol. 15, no. 2, 2000, pp. 42-47.
- [13] Balabanovic, M. and Y. Shoham, "Content-Based, Collaborative Recommendation," *Communications of the ACM*, vol. 40, no. 3, 1997, pp. 66-72.
- [14] Krulwich, B., "Lifestyle Finder," *AI Magazine*, vol. 18, no. 2, 1997, pp. 37-46.
- [15] de Vel, O. and S. Nesbitt, "A Collaborative Filtering Agent System for Dynamic Virtual Communities on the Web," *Working notes of Learning from Text and the Web, Conference on Automated Learning and Discovery CONALD-98*, Carnegie Mellon University, Pittsburgh, 1998.
- [16] Aridor, Y., et al., "Knowledge Agent on the Web," *Proceedings of the 4th International Workshop on Cooperative Information Agents IV*, 2000, pp. 15-26.
- [17] Chakrabarti, S., et al., "Focused Crawling: A New Approach to Topic-Specific Web Resource Discovery," *Proceedings of the Eighth International WWW Conference*, 1999, pp. 545-562.
- [18] Guarino, N., et al., "OntoSeek: content-based access to the Web," *IEEE Intelligent Systems*, vol. 14, no. 3, 1999, pp. 70-80.
- [19] Fensel, D., et al., "On2broker: Semantic-Based Access to Information Sources at the WWW," *Proceedings of the World Conference on the WWW and Internet (WebNet 99)*, Honolulu, Hawaii, USA, 1999, pp. 25-30.
- [20] Martin, P. and P. W. Eklund, "Knowledge retrieval and the World Wide Web," *IEEE Intelligent Systems*, vol. 15, no. 3, 2000, pp. 18-25.
- [21] Miller, G. A., "WordNet a lexical database for English," *Communications of the ACM*, vol. 38, no. 11, 1995, pp. 39-41.
- [22] Klein, D. A., *Decision-Analytic Intelligent Systems: Automated Explanation and Knowledge Acquisition*, Lawrence Erlbaum Associates, 1994.
- [23] Boose, J. H. 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, 1987, pp. 3-28.
- [24] Kleinberg, J. M., "Authoritative sources in a hyperlinked environment," *Journal of the ACM*, vol. 46, no. 5, 1999, pp. 604-632.
- [25] Kerschberg, L., et al., "A Semantic Taxonomy-Based Personalizable Meta-Search Agent," E-Center for E-Business, Working Paper, March, 2001.
- [26] Porter, M., "An Algorithm for Suffix Stripping," <http://www.muscat.co.uk/~martin/def.txt>.