Enhancing Web Search with Heterogeneous Semantic Knowledge

Rui Huang [1,2] and Zhongzhi Shi [1]

[1]Key Laboratory of Intelligent Information Processing

Institute of Computing Technology, Chinese Academy of Sciences

Beijing100190, China

[2]Graduate University of the Chinese Academy of Sciences

Beijing100049, China

huangr@ics.ict.ac.cn, shizz@ics.ict.ac.cn

Abstract: This paper explores four kinds of semantic knowledge to improve keyword-based Web search, including thesauruses, categories, ontologies, and social annotations. These heterogeneous semantic knowledge represent meanings of Web information, thus they can be used to improve search results in respect of semantic relevance. Currently, different semantic search paradigms have been developed for different kind of semantic knowledge respectively. However, how to make the most of all heterogeneous semantic knowledge to optimize Web search is still a big challenge in practice. To these ends, this paper proposes an integrated semantic search mechanism to incorporate textual information and keyword search with heterogeneous semantic knowledge and semantic search. Experiments show that the proposed mechanism effectively integrates heterogeneous semantic knowledge to improve Web search.

Keywords: Web search, semantic search, semantic Web, Web 2.0, ontology, social annotation

1. INTRODUCTION

Nowadays, search engines have been heavily relied on to retrieve information on the Web. Relevance ranking is vital to Web search paradigms, according to which potentially related Web documents with respect to the user's query are retrieved and ordered. As keyword-based Web search does not guarantee relevance in meanings, semantic search has recently attracted enormous and growing research focuses [1-5].

Heterogeneous semantic models have been introduced to represent the knowledge of interpreting the semantics of Web information and adopted for semantic search. Latent semantic models are induced from statistics of terms in documents, and used for semantic similarity computing [6]. Thesauruses (e.g. WordNet[1]) provide explanations of words and phrases as well as their synonyms and antonyms, which can be used for query expansion [7], similarity computing [8], etc. Categories (e.g. ODP[2]) include manually created classifications of Web documents according to their contents, with which to support category search. Ontologies [9] are manually formalized with commonly recognized knowledge in a certain domain, which can be used to understand data (semantic markups) on the Semantic Web [10]. They well support logical inference of semantic relations to obtain more exact semantic search results [11, 12], and can be combined to improve keyword-based search [2-4, 13]. Recent Web 2.0 [14] introduces social semantic annotations (e.g. social bookmarks on Del.icio.us[3]) assigned by common users to Web documents, which can be used to optimize search [15, 16, 5].

As all kinds of these semantic knowledge can help to interpret the meanings of Web information, to incorporate more of them would logically improve keyword search in respect of semantic relevance. However, to the best of our knowledge, no current search paradigm achieves such expected integration of heterogeneous semantic knowledge.

This paper proposes an integrated search mechanism to explore four kinds of semantic knowledge for keyword-based Web search, including thesauruses, categories, ontologies (and semantic markups), and social annotations. A statistical based measurement of semantic relevance, defined as semantic probabilities, is introduced to integrate heterogeneous semantic knowledge. It is calculated with all textual information and heterogeneous semantic knowledge, and stored in a newly proposed index structure called semantic-keyword dual index. Based on this uniform measurement, the search mechanism is developed to incorporate heterogeneous semantic knowledge for crawling, meta search and query expansion. Experiments show that the proposed mechanism can effectively integrate heterogeneous semantic knowledge to enhance Web Search in terms of semantic relevance.

Our work is among to first to improve Web search with heterogeneous semantic knowledge of all four mainstream semantic models. Two kinds of most related works are ontology and semantic markup based semantic search [2-4] and social annotation based search [15, 16, 5].

Mayfield et al. [2] propose to tightly integrate semantic inference and text retrieval, which is followed by works as [3] and [4] besides ours. [2] accepts keyword and semantic web queries separately, integrating only in retrieved results and through feedback mechanisms. Zhang et al. [3] use fuzzy description logic (DL) to integrate inference and retrieval, and thus accepts mainly formal DL queries. Tran

[1] http://wordnet.princeton.edu/
[2] http://dmoz.org/
[3] http://del.icio.us

et al. [4] present an ontology based approach to translate keyword queries to semantic web queries. Wu et al. [15] enlighten our statistical semantic relationship measurement, yet their work mainly focuses on social relationships of users. Dmitriev et al. [16] explore semantic relationships for enterprise search in which annotations are used as feedbacks. Bao et al. [5] improve Web search with social annotations and social page ranks. Our approach is designed for the open Web with different kinds of semantic knowledge. Statistical computing is adopted to integrate heterogeneous semantic knowledge throughout the whole process of crawling, indexing, query expansion and relevance ranking for Web search.

Rest of the paper is organized as follows. Section 2 describes the proposed semantic search mechanism, including overall framework, integration approach and search paradigm. Section 3 presents the experimental data and results and Section 4 concludes the paper.

2. SEMANTIC SEARCH MECHANISM

2.1 Overview

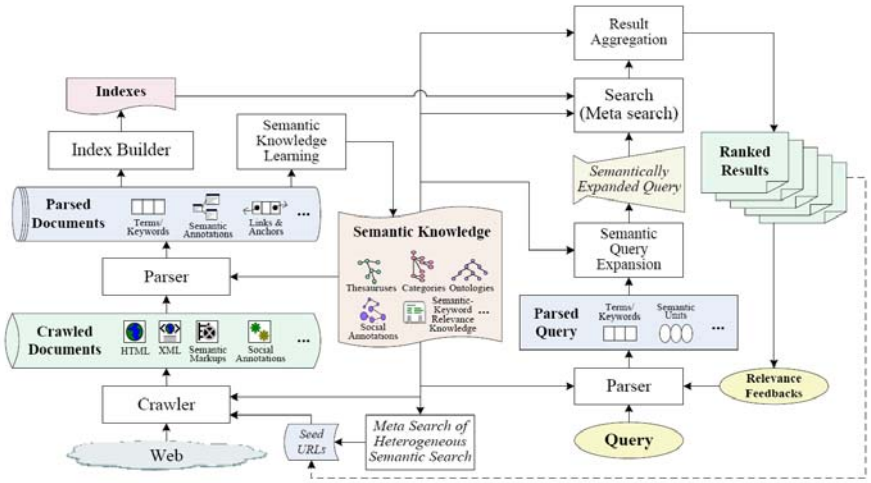

Fig. 1. Integrated Heterogeneous Semantic Search Framework

Fig. 1 illustrates the integrated heterogeneous semantic search framework. The knowledge base contains thesauruses, categories, ontologies, social annotations, and the automatically calculated semantic-keyword dual index (Section 2.2.3).

Based on semantic knowledge, related documents are crawled from the Web and parsed for the document corpus. The crawler and the parser can also use such domain knowledge for topical crawling (Section 2.3.1). The parsed document base includes not only keywords and links, but also semantic annotations (e.g. semantic markups, social annotations). When a new query is issued or relevance feedbacks are given, they are parsed and expanded with both keywords and heterogeneous semantic knowledge in the knowledge base (Section 2.3.2). Search (or meta-search) results are ranked and aggregated based on the semantic knowledge base and presented for the user (Section 2.3.3 & 2.3.4). For those URLs which are relevant to the query, yet not included in the crawling list, crawl them to update the document corpus, semantic knowledge base and index base.

2.2 Integration of heterogeneous semantic knowledge

2.2.1 Definition: semantic probabilities

Heterogeneous semantic knowledge are integrated with a uniform measure of semantic probabilities. Since not all semantic models involved are easily formalized (e.g. social annotations), we propose to adopt statistical computing in completion, and recast the definitions of traditional probabilities to represent both keyword based textual information and heterogeneous semantic knowledge.

Each keyword based textual (Web) document and its corresponding semantic knowledge is represented with a semantic annotation $sa = [T, S]$, where T is the list of all keyword based terms in the document, and S is the list of all semantic unit in the semantic knowledge. Here, semantic unit defines the minimum unit that represents certain complete and clear semantics, such as a phrase in the thesaurus, a category name, a concept in the ontology, or a social annotation.

In a semantic annotation $sa = [T, S]$, $\forall t_i \in T$ is called semantically occur in sa, represented as $t_i \vdash_{sa}^{T}$. $\forall s_j \in S$ is called semantically occur in sa, represented as $s_j \vdash_{sa}^{S}$. If $t_i \vdash_{sa}^{T} \wedge s_j \vdash_{sa}^{S}$, then t_i and s_j are called semantically cooccur in sa, represented as $< t_i, s_j > \vdash_{sa}$.

All Web documents and semantic information construct the semantic annotation space $SA = (sa_1, ..., sa_n)$, in which semantic occurrence probability, semantic cooccurrence probability and semantic conditional probability are defined.

Definition 1 (Semantic Occurrence Probability)

$\forall(t_i \vdash_{sa}^{T}) \wedge (sa \in SA)$, $P(t_i) \in [0,1]$ denotes the semantic occurrence probability that

term t_i semantically occurs in SA. $\forall(s_j \vdash_{sa}^{S}) \wedge (sa \in SA)$, $P(s_j) \in [0,1]$ denotes the semantic occurrence probability that semantic unit s_j semantically occurs in SA.

Definition 2 (Semantic Cooccurrence Probability)

$\forall(< t_i, s_j > \vdash_{sa}) \wedge (sa \in SA)$, $P(< t_i, s_j >) \in [0,1]$ denotes the probability that term t_i semantically cooccurs with semantic unit s_j in SA.

Definition 3 (Semantic Conditional Probability)

$P(t_i / s_j) = P(< t_i, s_j >) / P(s_j) \in [0,1]$ denotes the semantic conditional probability of term t_i given semantic unit s_j, $P(s_j / t_i) = P(< t_i, s_j >) / P(t_i) \in [0,1]$ denotes the semantic conditional probability of semantic unit s_j given term t_i.

2.2.2 Computation model

In order to compute the above defined semantic probabilities, traditional TF-IDF computation model is extended to cover both terms and semantic units, represented with formulae (1)-(4).

$$P(t_i) = (\sum_{sa=[T,S] \in SA, t_i \vdash_{sa}^{T}} (\ln \frac{\|SA\| - df(t_i) + 0.5}{df(t_i) + 0.5} * \frac{1 + \ln(1 + \ln tf(t_i, T))}{(1-\gamma) + \gamma \frac{\|T\|}{avTl}})) / \|SA\| \quad (1)$$

$$P(s_j) = (\sum_{sa=[T,S] \in SA, s_j \vdash_{sa}^{S}} (\ln \frac{\|SA\| - df(s_j) + 0.5}{df(s_j) + 0.5} * \frac{1 + \ln(1 + \ln tf(s_j, S))}{(1-\delta) + \delta \frac{\|S\|}{avSl}})) / \|SA\| \quad (2)$$

$$P(< t_i, s_j >) = (\sum_{sa=[T,S] \in SA, t_i \vdash_{sa}^{T} \wedge s_j \vdash_{sa}^{S}} (\ln \frac{\|SA\| - df(t_i) + 0.5}{df(t_i) + 0.5} * \ln \frac{\|SA\| - df(s_j) + 0.5}{df(s_j) + 0.5} * \frac{1 + \ln(1 + \ln tf(t_i, T))}{(1-\gamma) + \gamma \frac{\|T\|}{avTl}} * \frac{1 + \ln(1 + \ln tf(s_j, S))}{(1-\delta) + \delta \frac{\|S\|}{avSl}})) / \|SA\| \quad (3)$$

$$P(t_i / s_j) = P(< t_i, s_j >) / P(s_j),\ P(s_j / t_i) = P(< t_i, s_j >) / P(t_i) \quad (4)$$

where $df(t_i)$ is the document frequency of term t_i in the semantic space SA, $tf(t_i, T)$ is the term frequency of t_i in a semantic annotation $sa = [T, S]$, $df(s_j)$ is the document frequency of semantic unit s_j in SA, $tf(s_j, S)$ is the term frequency of s_j in sa. *avTl* is the average length of the list of terms, *avSl* is the average length of the list of semantic units, γ and δ are used to balance the length.

The semantic occurrence probability of each term (or semantic unit) is calculated as the arithmetic average of the TF-IDF of the term (or semantic unit) in each semantic annotation in the semantic space. The semantic cooccurrence probability of a term and a semantic unit is calculated as the arithmetic average of the product of TF-IDF of the term and that of the semantic unit in each semantic annotation in the semantic space. The semantic conditional probability is calculated with semantic occurrence and cooccurrence probability according to its definition.

2.2.3 Storage: semantic-keyword dual index

To effectively retrieve the semantic relevance between all terms and semantic units, semantic-keyword dual index structure is proposed which consist of a pair of interrelated indexes: a term-semantic inverted index and a semantic-term inverted index (shown in Fig. 2).

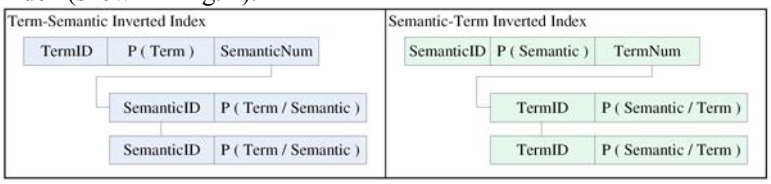

Fig. 2. Semantic-keyword dual index structure

Term-semantic inverted index is ordered by *TermID* unique to each term. $P(Term)$ is the semantic occurrence probabilities of Term. *SemanticNum* stores the length of SemanticID links, indicating the total number of related semantic units. $P(Term / Semantic)$ is the semantic conditional probability of the term *Term* given the semantic unit *Semantic*. Similar structure holds for semantic-term inverted index. The two interrelated inverted indexes compose the semantic-keyword dual index of statistical semantic relevance knowledge.

2.3 Search with heterogeneous semantic knowledge

2.3.1 Semantic knowledge based crawling

At the beginning of crawling, domain specific semantic knowledge are specified, which are used for meta-search to find seed URLs for both textual documents and semantic information.

A piece of crawled Web document (e.g. an XML document or an HTML webpage) along with its related semantic information (e.g. ontology based semantic markups or semantic annotations) is represented as a semantic annotation $sa_{crawled} = [T_{crawled}, S_{crawled}]$. The specified domain is represented as the whole semantic space SA. Then the semantic relevance of the crawled semantic annotation with the specified domain is represented as $SR(sa_{crawled} / SA)$.

$$SR(sa_{crawled} / SA = (sa_1, ..., sa_p)) = \left(\sum_{i=1}^{\|SA\|} SR(sa_{crawled} / sa_i)^2 \Big/ \|SA\| - 1 \right)^{\frac{1}{2}} \quad (5)$$

where $SR(sa_{crawled}/sa_i) =$

$$\frac{\sum_{tcj \in Tsa_{crawled}} \sum_{tj \in Tsa_i} \sum_{sck \in Ssa_{crawled}} \sum_{sk \in Ssa_i} Max(P(t_{cj}/t_j), P(s_{ck}/t_j), P(t_{cj}/s_k), P(s_{ck}/s_k))}{\|Tsa_{new}\| \cdot \|Tsa_i\| \cdot \|Ssa_{crawled}\| \cdot \|Ssa_i\|} \quad (6)$$

As formulae (5)-(6) shows, the similarity of the crawled $sa_{crawled}$ with the specified domain (semantic annotation space SA) is the geometric average of the similarities of $sa_{crawled}$ with each semantic annotation sa_i in SA. The semantic similarity of $sa_{crawled}$ with sa_i is the arithmetic average of the max conditional probability of all terms and semantic units in $sa_{crawled}$ given all terms and semantic unites in sa_i. If the $SR(sa_{new}/SA)$ falls above a minimum relevance threshold, then the document and semantic information is considered relevant to the specified domain, and will be parsed and indexed.

2.3.2 Semantic query expansion

To semantically expand the queries based on statistical semantic similarities among both keywords ant semantic units, three steps of semantic computing, semantic inference and query expansion are involved.

In semantic computing, the semantic-keyword dual index is iteratively searched to find all possible keywords and semantic units of the user specified query (either in the form of traditional keyword query, semantic web query or combination of both). This process of iterative search adopts theories of spread activation [17]. It starts with the original query. For each keyword and semantic unit or, newly found keyword and semantic unit, find its related keywords and semantic units along with their relevance. Termination control is achieved with a decay factor (weighted less after each iteration) and a minimum threshold (terminate when semantic probability is below the threshold). Semantic inference uses ontologies in the semantic knowledge base to infer more semantically related semantic units. Weight of the inferred semantic unit is calculated as the product of the weight of inference and the weight of original semantic unit. Query expansion includes all computed and inferred keywords and semantic unites to expand the original query.

2.3.3 Semantic relevance ranking

Let T be the set of all terms and S be the set of all semantic units. A document and semantic information is represented as $D = [d_T \quad d_S]^T$, where d_T is the term frequency of all terms in *D*, and d_S is the term frequency of all semantic units in *D*. The expanded query is represented as $Q' = [q_T \quad q_S]^T$, where q_T is the calculated weight of all terms, and q_S is the calculated weight of all semantic units.

Similarity of D w.r.t. Q' in T and S is represented as $Sim(Q'/D)\big|_{T\cup S}$, which can be calculated as the arithmetic average of the semantic conditional probabilities of each keyword or semantic unit in Q' given each keyword or semantic unit in D (shown in formula (7)).

$$Sim(Q'/D)\big|_{T\cup S} = \sum_{i=1}^{\|T\cup S\|}\sum_{j=1}^{\|T\cup S\|} P(q_i/d_j) \Big/ \|T\cup S\| \cdot \|T\cup S\| \tag{7}$$

2.3.4 Meta search of heterogeneous search engines

The proposed semantic search mechanism also supports meta-search of current heterogeneous search engines. If a domain is specified for the search engine (so called topical search or vertical search engine), then meta-search can be employed to obtain better related seed URLs.

Top search results of each search engine are included (top 500 in this paper). To integrate meta-search results, joint relevance scores are computed according to formula (7). Rank based merge does not fit for heterogeneous semantic search that adopt quite different ranking criteria respectively. The joint relevance ranking take into consideration possible semantic relevance among all keywords and semantic units, even if different result documents are found in different search engines.

3. RESULTS AND DISCUSSIONS

3.1 Datasets

The WordNet thesaurus[4] is used both for stemming and as prior knowledge. ODP data of 730,416 categories obtained May, 2007[5] are included. Three domains of computer, sports and entertainment are experimented. For semantic web data, we use the 10,429,951 RDF triples extracted from Swoogle cache June, 2005[6] and 347 ontologies crawled from the Web. We also crawled a sample of Del.icio.us data during May, 2007, consisting of 459,143 social annotations covering 28,704 different links, 184,136 different users and 54,460 different tags.

---

4. http://wordnet.princeton.edu/obtain
5. http://rdf.dmoz.org/rdf/
6. http://ebiquity.umbc.edu/resource/html/id/126/10M-RDF-triples

3.2 Preliminary Results

With heterogeneous semantic knowledge, queries are expanded with both keywords and semantic units. For the query "semantic web", the top 10 related keywords are web, semantic, rdf, ontology, xml, w3c, research, data, owl, and knowledge. The top 10 related semantic units are *semantic*, *rdf*, *web*, *reference*, *category:topic*, *ontology*, *owl*, *rdfs:comment*, *web2.0*, and *srwc:isworkedonby*.

For each domain of computer, sports and entertainment, we specify an ontology and a wordlist of related keywords. Then related documents are crawled, parsed and indexed. Breath first strategy is used by the crawler. Results in Table 1 show that the crawlers are effective in topical crawling.

Table 1. Semantic knowledge based crawling statistics

Topic	Visited URLs	Analyzed URLs	Related URLs	Harvest Rate
Computer	16,703	232,991	11,965	71.63%
Sports	19,996	399,110	14,208	71.05%
Entertainment	14,375	303,744	8,964	62.36%

To test the effectiveness of semantic relevance ranking, 3000 web pages are crawled and selected from the corresponding category in Yahoo for each topic to see whether our relevance ranking algorithm takes it as semantically relevant. Table 2 proves the correctness of our proposed algorithm.

Table 2. Semantic relevance ranking correctness

Topic	Related in Yahoo	Related in our algorithm	Correctness
Computer	3000	2,654	88.48%
Sports	3000	2,955	98.50%
Entertainment	3000	2,791	93.03%

4. CONCLUSIONS

This paper proposes to improve Web search with both keyword-based textual information and heterogeneous semantic knowledge. A uniform statistical semantic measure along with its computation model and storage structure is proposed to represent semantic relevance of all keywords and heterogeneous semantic models. Based on this uniform measure, the proposed mechanism well supports semantic knowledge based crawling, query expansion, relevance ranking and meta search. Experimental results prove the effectiveness of the proposed mechanism.

In the future, we will focus on evaluation and optimization of semantic crawling, relevance ranking and heterogeneous result aggregation algorithms. Moreover, we will improve the system for public use.

ACKNOWLEDGEMENTS

This work is supported by the 973 National Basic Research Programme (No.2007CB311004), the 863 National High-Tech Program (No.2006AA01Z128), and the National Natural Science Foundation of China (No.90604017, No.60435010, No.60775035).

REFERENCES

1. Guha R., Mccool, R., and Miller. E. “Semantic search”, In Proceedings of WWW '03, pp.700-709, 2003.
2. Mayfield, J., and Finin T. “Information retrieval on the semantic web: Integrating inference and retrieval”, In SIGIR 2003 Semantic Web Workshop, 2003.
3. Zhang, L., Yu, Y., Zhou, J., Lin, C., and Yang, Y., “An enhanced model for searching in semantic portals”, In Proceedings of WWW '05, pp.453-462, 2005.
4. Tran, T., Cimiano, P., Rudolph, S., and Studer, R., “Ontology-Based Interpretation of Keywords for Semantic Search”, In Proceedings of ISWC '07, pp.523–536, 2007.
5. Bao, S., Wu, X., Fei, B., Xue, G., Su, Z., and Yu, Y., “Optimizing web search using social annotations”, In Proceedings of WWW '07, pp.501-510, 2007.
6. Furnas, G.W., Deerwester, S., Dumais, S.T., Landauer, T.K., Harshman, R.A., Streeter, L.A., and Lochbaum, K.E., “Information retrieval using a singular value decomposition model of latent semantic structure”, In Proceedings of SIGIR '88, pp.465-480, 1988.
7. Voorhees, E.M., “Query expansion using lexical semantic relations”, In Proceedings of SIGIR '94, pp.61-69, 1994.
8. Tollari, S., Glotin, H., and Maitre, J.L., “Enhancement of textual images classification using segmented visual contents for image search engine”, Multimedia Tools and Applications, vol.25, No.3, pp.405-417, 2005.
9. Studer, R., Benjamins, V.R., and Fensel, D., “Knowledge engineering: principles and methods”, Data and Knowledge Engineering, vol.25, No.1-2, pp.161-197, 1998.
10. Berners-Lee, T., Hendler, J., and Lassila, O., “The semantic web”, Scientific American, vol.284, No.5, pp.34-43, 2001.
11. Cohen, S., Mamou, J., Kanza, Y., and Sagiv, Y., “Xsearch: A semantic search engine for xml”, In Proceedings of VLDB '03, pp.45-56, 2003.
12. Ding, L., Finin, T., Joshi, A., Peng, Y., Pan, R., and Reddivari, P., “Search on the semantic web”, IEEE Computer, vol.10, No.38, pp.62-69, 2005.
13. Rocha, C., Schwabe, D., and de Aragao, M.P., “A hybrid approach for searching in the semantic web”, In Proceedings of WWW '04, pp.374-383, 2004.
14. O'Reilly, T., “What is web 2.0: Design patterns and business models for the next generation of software”, O'Reilly (http://www.oreilly.com/), September 2005.
15. Wu, X., Zhang, L., and Yu, Y., “Exploring social annotations for the semantic web”, In Proceedings of WWW '06, pp.417-426, 2006.
16. Dmitriev, D.A., Eiron, N., Fontoura, M., and Shekita, E., “Using annotations in enterprise search”, In Proceedings of WWW '06, pp.811-817, 2006.
17. Crestani, F., “Application of spreading activation techniques in information retrieval” Artificial Intelligence Review, vol.11, No.6, pp.453-482, 1997.