Adjusting to Specialties of Search Engines Using MetaWeaver

Mikihiko Mori and Seiji Yamada
CISS, IGSSE, Tokyo Institute of Technology
4259 Nagatsuta, Midori, Yokohama 226-8502, Japan
mori@ymd.dis.titech.ac.jp

Abstract:

In this paper, we propose MetaWeaver which can selectively utilize effective ones from registered search engines depending on queries. Traditional meta-search engines use all the registered search engines without considering which search engines are appropriate, and often return many irrelevant Web pages in the hit list. In contrast with that, MetaWeaver is able to evaluate the specialties of each search engine by analyzing the hit list for sample queries, and to utilize them for selecting adequate search engines for current queries. Also we found out our approach is promising by making preliminary experiments.

1 Introduction

There are currently a lot of search engines to retrieve relevant Web pages to a query from the World Wide Web (WWW). In view of the characteristics of the stored Web pages, the search engines are categorized into two groups. One group is general all-round search engines like AltaVista[1]. The other group is special search engines like DogInfo[2], which is a dog-specific search engine. Because a general search engine have a huge database, we use it for searching any query. However there is a significant issue that the coverage of a general search engine is significantly insufficient. The Web size and the size of each search engine's database were surveyed in (Bharat & Broder 1998). The largest search engines covered 50% of all of the Web pages and maximum overlap of search engines is 30%. Thus a user can not search well using a single general search engine.

A meta-search engine solves the issue by ask other search engines about the queries. It does not collect pages in the Web by itself. Instead it registers other search engines in advance, and requests them to retrieve for the queries and gather their results. Then the meta-search engine provides a hit list of Web pages by integrating the hit lists from other search engines and removing overlapping pages. However a meta-search engine also has a significant problem: it can not selectively utilize registered search engines depending on their specialties. Hence it always requests all the registered search engines and the integrated hit list include many irrelevant Web pages. Though some meta-search engine ask a user to select adequate registered search engines, he/she hardly recognize specialties of each search engine on the queries and select effective ones from them.

We propose MetaWeaver which can grasps specialties of each search engine and request search engines whose specialty is content of the queries. When it estimates search engines, it uses thesaurus for improvement of estimate. Therefore, it is expected that Web pages from it satisfies user's queries more precisely than traditional meta-search engines.

MetaCrawler (Selberg & Etzioni 1997) requests all handled search engines in parallel and removes duplication. Inquirus (Lawrence & Giles 1998) also submits all search engines. However, it downloads the Web pages in hit lists from the search engines and then, it calculates relevance to the query using downloaded pages.

On the other hand, ProFusion (Fan & Gauch 1999) selects appropriate search engines to request them to search. It manually builds categories from the words of the newsgroups names and it can only select when the words include words of user's query.

SavvySearch (Howe & Dreilinger 1997) can also select proper search engines. It uses words of users' queries and their effectiveness values. Visited links by users evaluate the value. SavvySearch, however, does not use any thesaurus. Therefore, it will not evaluate search engines on initial steps of searches.

[1] http://www.altavista.com/

[2] http://www.doginfo.com/

2 MetaWeaver

Meta-search engines need to choose proper search engines for queries a user requested. MetaWeaver can detect effective search engines by the word used as in queries. The architecture is shown in Figure 1.

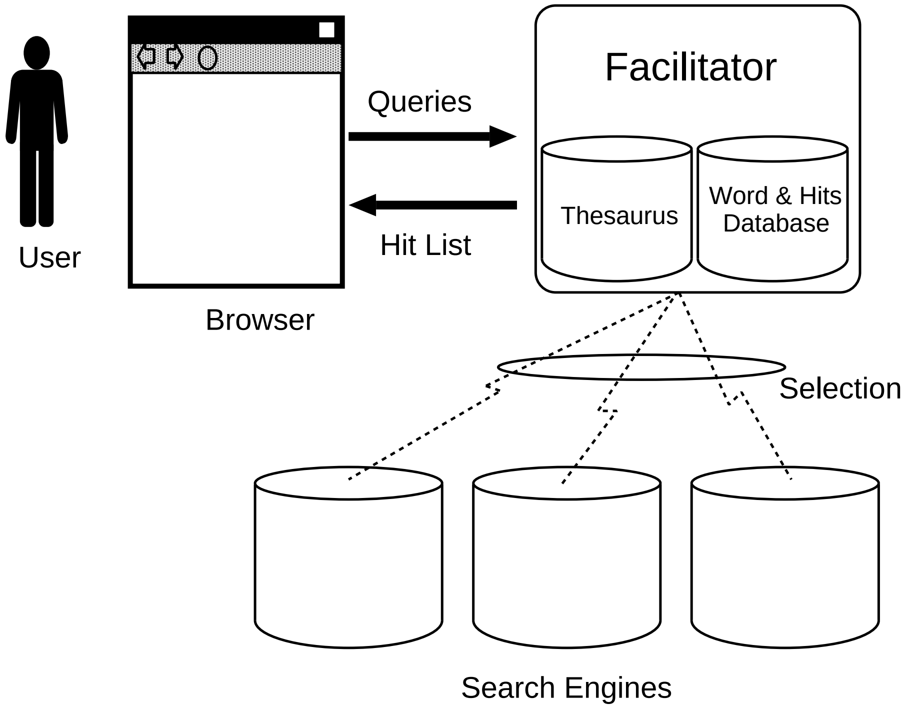

Figure 1: The architecture of MetaWeaver

2.1 Search Overview

A user writes queries to the meta-search engine to submit them on a his/her browser. Users can make queries as AND, OR search or these combinations.

The facilitator selects some search engines from registered search engines as appropriate search engines of the user's query when a user make a submission. We define a suitability to choose appropriate search engines and the detail of suitability is described in section 3.

The facilitator send user's query to selected engines in parallel. It waits while all engines return lists of top 10 hits. When it gets the hit lists, it integrates into a hit list to display the user's browser. The way of integration for making the meta-search hit list is described in section 4.

The suitability is updated after the user gets the hit list. The suitability of each word in the query is based on the number of hits of the word. Therefore, the facilitator gets the number of hits of each word from each of all registered search engines and writes the numbers into the word and hits database with search engine names. Thus, MetaWeaver behaves adaptively toward the changes of search engines' databases.

2.2 System Requirements

MetaWeaver utilizes AltaVista[3], Yahoo[4] and Infoseek[5] as general search engines and also utilizes DogInfo[6] (a dog specific search engine), BioCrawler[7] (a biological search engine) and LawCrawler[8] (a law search engine) as specialty search engines.

For thesaurus, MetaWeaver utilizes the EDR Electronic Dictionary[9]that contains the relation between four hundred thousand concepts.

3 Calculation of The Suitability

MetaWeaver needs to measure the suitabilities for choosing appropriate search engines. The appropriate search engines is regarded as search engines which returned large number of hits according to queries before. Larger hits tend to be less precision, which is the relevant pages per hit pages, but more relevant pages actually include.

In this section, we present a method to measure the suitability of a single-word and a combined-word query.

3.1 A Single-word query

When a query consists of a single word, the suitability $p_{s,w}$ of a search engine s for word w is the number of hits divided by the size of search engine s:

$$p_{s,w} = \frac{u_{s,w}}{U_s} \tag{1}$$

where $u_{s,w}$ is the number of hits from the word and hits database and U_s is the search engine's size.

However, it is not usual that the search engine opens its size. Therefore, MetaWeaver uses the history of hits from the word and hits database for the last f_s words of each search engine. f_s is currently 100. We express U_s instead of exact search engine size as follows:

$$U_s = \alpha_u \sum_{i=1}^{f_s} u_{s,w_{n-i}} \tag{2}$$

where u_{w_n} is the number of hits of the nth word from the last and α_u is the correction factor for approximating summation of u_{s,w_n} to U_s.

Here, the thesaurus improves $p_{s,w}$. If MetaWeaver does not use any thesaurus on initial steps of searches, it does not initially decide right estimation of search engines for choosing appropriate ones because it estimates the suitability at zero not to have the number of hits about the words in a query in the word and hits database.

By using thesaurus, MetaWeaver to be able to compute $p_{s,w}$ not only when w do not have any records in the word and hits database but also when w have records. The suitability after improvement by thesaurus is described:

$$p_{s,w}^{imp} = p_{s,w} + \frac{\sum_{m \in W_t} \delta_{w,m} p_{s,w}}{|W_t|} \tag{3}$$

where $p_{s,w}^{imp}$ is the improved suitability, $p_{s,w}$ is the suitability of each word in a synonym set W_t and $\delta_{w,m}$ is the weight of relation between w and a synonym m. $\delta_{w,m}$ between each word and each of its synonyms is set 1 currently.

3.2 A Combined-Word Query

We consider the case in which a query contains several words combined with boolean AND / OR. The result of AND searches are generally smaller than that of each word in the queries, and that of OR searches are oppositely larger than that of each word. Hence we express the suitability of an AND / OR search as follows.

[3] http://www.altavista.com/

[4] http://www.yahoo.com/

[5] http://infoseek.go.com/

[6] http://www.doginfo.com/

[7] http://www.biocrawler.com/

[8] http://lawcrawler.findlaw.com/

[9] http://www.iijnet.or.jp/edr/index.html

AND search The suitability is calculated as the geometric mean of every $p_{s,w}^{imp}$ for every w in the query q:

$$p_{s,q} = \left(\prod_{w \in q} p_{s,w}^{imp} \right)^{\frac{1}{|q|}} \tag{4}$$

OR search The suitability is calculated as the average value of every $p_{s,w}^{imp}$ for every w in the query q:

$$p_{s,q} = \frac{\sum_{w \in q} p_{s,w}^{imp}}{|q|} \tag{5}$$

4 Ranking Web Pages from the Search Engines

For obtaining a single hit list, MetaWeaver ranks Web pages in hit lists to order them. MetaWeaver employs the suitability of each search engine and hit ranking of each Web pages in the hit list from each engine for the single list.

A score $\sigma(p)$ of a Web page p expresses as follows:

$$\sigma(p) = \sum_{s \in S^{ap}} \frac{p_{s,q}}{k(p)} \tag{6}$$

where S^{ap} is appropriate (query-requested) search engines, $k(p)$ is the ranking of the page p in a hit list from a search engine.

Here, the top page of the list obtains 1 and the 10th page obtains 10. If a search engine s does not have p in the list, $\frac{p_{s,q}}{k(p)}$ is zero.

5 Experiments

We made experiments to evaluate the effectiveness of using suitability in MetaWeaver. To simplify the experiments, we do not employ thesaurus for the improvement of the suitability, and the word and hits database has already stored randomly 2000 words in the thesaurus and their hits. Also the queries for a evaluation in the experiment are ‘dna’, ‘soccer’ as single-word queries, ‘golden AND retriever’ and ‘punitive AND law’ as combined-word queries. We choose 100 for the correction factor f_s that is the history of hits from the word and hits database and 10 for α_u.

Experimental results are shown in Table 1. In the table, “IS” means Infoseek, “AV” means AltaVista, “YH” means Yahoo, “BC” means BioCrawler, “LC” means LawCrawler and “DI”means DogInfo. If MetaWeaver selects correctly appropriate search engines, it will get right lists of ranking of search engines for each query. The lists are sorted largest first by the suitabilities of the search engines.

The results indicate our aiming tendency toward estimating specialty search engines’ suitabilities at the top. The query ‘soccer’ does not have precise specialty search engines. Thus, no specialty search engine is the top and the general search engines occupy the top three. The suitability of the top one or two is ten times larger than that of below.

Table 1: The search engine ranking of the queries

	dna		soccer		golden retriever		punitive law	
Ranking	SE	p_w	SE	p_w	SE	p_w	SE	p_w
1	BC	0.02336	AV	0.01317	DI	0.11111	LC	0.00203
2	DI	0.01296	IS	0.00311	YH	0.00026	AV	0.00015
3	LC	0.00715	YH	0.00285	AV	0.00018	YH	0.00000
4	AV	0.00346	LC	0.00148	IS	0.00003	IS	0.00000
5	IS	0.00187	DI	0.00092	LC	0.00000	DI	0.00000
6	YH	0.00047	BC	0.00000	BC	0.00000	BC	0.00000

6 Conclusion

Ordinary meta-search engines present too abundant Web pages over since they had submitted all of the registered search engines. Also specialty-sensitive meta-search engines like SavvySearch does not consider effect of thesaurus. MetaWeaver solves both disadvantages. MetaWeaver selectively utilizes fewer search engines with a suitability.

In the experimental results, we confirm that the definition of the suitability is valid. We will use this system for a long time to investigate whether it adapt real person's queries or not.

References

Howe, A.E. and Dreilinger, D. (1997). *SavvySearch: A Meta-search Engine that Learns Which Search Engines to Query*, AI Magazine, 18 (2), 19–25.

Selberg, E. and Etzioni, O. (1997). *The MetaCrawler Architecture for Resource Aggregation on the Web in both Postscript and HTML*, IEEE Expert, 12 (1), 8–14.

Bharat, K. and Broder, A. (1998). *A technique for measuring the relative size and overlap of Web search engines*, Proceedings of the 7th International World Wide Web Conference.

Lawrence, S. and Giles, C.L. (1998). *Inquirus, the NECI meta search engine*, Proceedings of the 7th International World Wide Web Conference.

Fan, Y. and Gauch, S. (1999). *Adaptive Agents for Information Gathering from Multiple, Distributed Information Source*, Proceedings of 1999 AAAI Spring Symposium on Intelligent Agents in Cyberspace, 40–46.