# Experiences in Building a High Quality Meta Search Engine
## –Poster Submission–

Julia Böttcher Felix Naumann
Humboldt-Universität zu Berlin
{boettche|naumann}@dbis.informatik.hu-berlin.de

## 1 High Quality Information Integration

The development of the Internet—especially the World Wide Web—has made it possible and useful to access a multitude of information sources on almost any given topic. Web directories guide users towards these sources, search engines let users discover sources previously unknown, and a huge number of web sites act as information sources and provide the actual information.

Most often a user can choose between many alternative information sources to obtain the desired information. Consider the numerous search engines available on the web. Most users have chosen their favorite search engine, possibly based on personal experience in response time, relevancy of the results, completeness, etc. However, users might miss just the right web page to their query, simply because that page was not (yet) indexed by that search engine. Meanwhile, this web page might have already been indexed by other search engines. The user might turn to one of the others and may eventually find the desired link. A meta search engine (MSE) solves this problem by simultaneously querying multiple search engines with the users query. The results of the different engines are integrated to a combined response to the user. The drawback of this simple approach is that the MSE possibly uses search engines that are deemed qualitatively bad by the user. A MSE might access sources with low response time, outdated links, unreasonable or no document ranking, etc. Quality reasoning about the search engines helps identify good sources in a user dependent and query dependent way.

Additionally, quality reasoning can contribute to high quality integration of information: First, the results of search engines can be ranked according to their individual quality. We both integrate ranking techniques of the search engines proper [GGM97], and use a new quality-based ranking to integrate and rank the multiple search results. Second, information conflicts can be resolved better if the quality of the conflicting values is known. Due to the fleeting nature of the web and due to differing crawling techniques, search engines might return differing data about the same web page. Depending on the type of conflict, quality measures can help resolve them, for instance by favoring search engines with more frequent updates. The following sections describe our High Quality Information Querying (HiQIQ) approach to an implementation of a meta search engine.

## 2 Information Quality on the Web

WWW information sources display large differences in the quality of the information they present. The information can be up-to-date or outdated, accurate or inaccurate, costly or free, fast or slow, comprehensible or unclear, complete or incomplete, etc. Experienced users of the web will have come across sources to which an any combination of these and more adjectives can be applied. Most often information quality (IQ) is not as high as one could wish or would expect. Search engines are no different. In particular for search engines, we can identify the eight IQ criteria of Table 1. For more detailed descriptions of the criteria see [NR00].

| IQ Criterion | Description |
|---|---|
| Accuracy | Quality of the result ordering |
| Age | Update frequency of a search engine |
| Availability | Percentage of time an information source is "up" |
| Coverage | Percentage of the web that a search engine has indexed |
| Density | Number of attributes a search engine exports |
| Latency | Time until the first web link reaches user |
| Redundancy | Number of "unnecessary" links in a search result |
| Response time | Time until the complete response reaches the user |

Table 1: IQ criteria for WWW search engines

To apply IQ reasoning techniques with the help of IQ criteria we need three components: (i) IQ criteria assessment methods determine scores for each criterion. For some criteria, the methods are automatic, others need user input [NR00]. (ii) Ranking methods like the SAW method use the multiple IQ criteria scores to find a weighted ranking [Nau98]. (iii) Optimization algorithms that determine good (or optimal) combinations of sources to query [NLF99]. Put together, these components build the core source selection facility for any information integrating service, and in particular for our HiQIQ meta search engine.

# 3 Implementation of the HiQIQ Meta Search Engine

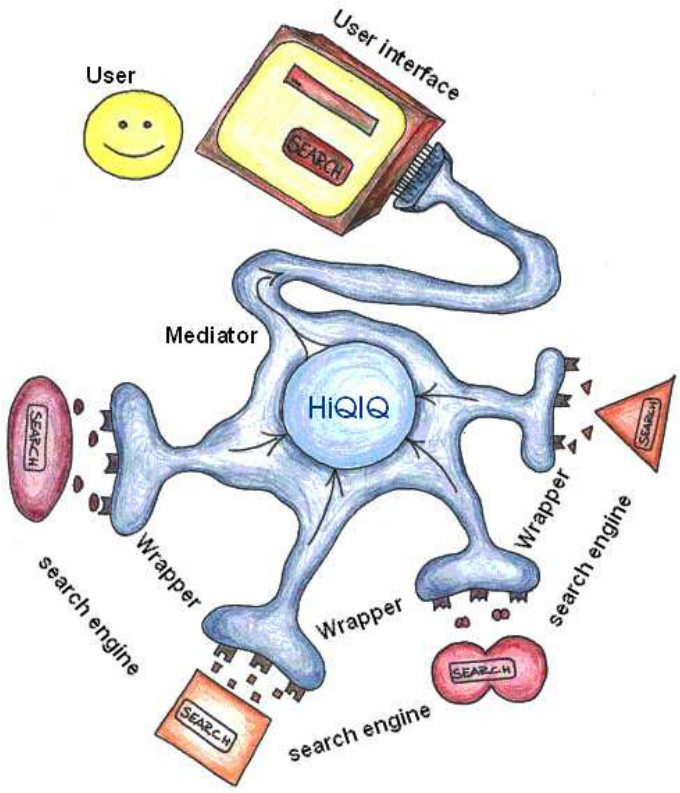


Figure 1: The Mediator-Wrapper Architecture

For the implementation of our meta search engine we followed the mediator-wrapper architecture as proposed by Wiederhold (Figure 1, [Wie92]). The user poses a query (a set of keywords) through a user interface against the global schema of the mediator. For search engines, the global schema is but one relation with the URL as ID and the attributes name, description, size, date, language, rating, and ranking. The mediator, in turn, selects appropriate search engines and sends the queries to the corresponding wrappers. These wrappers translate the query into `http` requests that are understood by the search engine they wrap. Their response is again translated by each wrapper into Java objects containing XML elements and sent to the mediator. There, the mediator integrates the information to a single high quality response to the user. The entire HiQIQ meta search engine is implemented in Java.

We are aware of existing meta search engines like MetaCrawler or SavvySearch, but often find their results unsatisfying. First, the user has no influence on search engine selection. Second, none integrate the attribute values of all search engines, i.e., their results are somewhat sparse. Third, as far as we know, none of the search engines considers quality aspects such as the number of dead links, availability, or accuracy in their source selection and ranking methods.

**Wrapper.** The task of a wrapper component is to query a certain search engine and to translate its answer into a determined data format. We have implemented one wrapper for each engine that serves as source for the meta search Engine. A wrapper gets the search keywords from the mediator and communicates with its search engine via `http` (Figure 2). As soon as an answer returns, the wrapper extracts the essential information and passes it back to the mediator as a Java object (Figure 3). If necessary, the wrapper accesses the search engine multiple times.

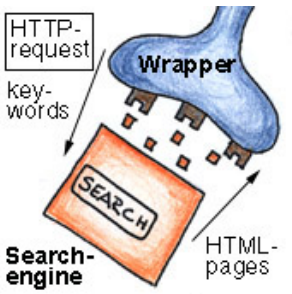

Figure 2: Search engine

For building the wrappers we used the World Wide Web Wrapper Factory (W4F) which provides an easy extraction language for specifying the source that shall be wrapped [SA99]. Specifying a wrapper takes an average of five hours work. Currently we have wrapped AltaVista, Excite, Fast, Go, Google, Hotbot, Northern Light, and Voila. Out goal is to include as many search engines as possible. Due to our HiQIQ approach we do not have to worry about search engines being small, slow, or having a low update frequency. We recognize low quality and only access such sources when necessary.

As other authors have pointed out [Coh99] and as we have observed as well, a problem not to be underestimated is to keep the meta search engine working properly. The output of the different search engines changes from time to time and therefore the wrappers have to be updated frequently.

**Mediator.** The Mediator is the heart of the HiQIQ meta search engine. It selects appropriate search engines by using quality measures and calls the corresponding wrappers. Their results are collected, merged, and ranked with respect to the determined quality of the sources. The mediator is also connected with the user interface, in order to receive the search request and to deliver the combination of all results to it.

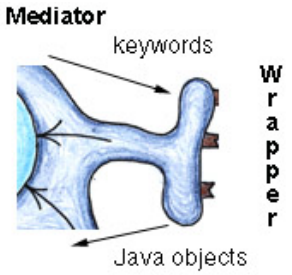

Figure 3: Wrapper

For performance reasons the wrappers are called in parallel as threads. This design decision was made due to the fact that the actual quality calculations of wrappers take very little time compared with the time that is needed for querying all the search engines.

Besides collecting the answers of the wrappers, the mediator is also responsible for integration and ranking of the search results. If two results with the same URL are received, the data is merged and treated as a single result. So called resolution functions are used to resolve conflicts between values. For the moment we resolve conflicts by concatenating the conflicting values, in the future we plan to use quality measures to decide, which search engine "wins" the conflict (Figure 6).

**User interface.** The graphical user interface of the HiQIQ meta search engine is modeled according to the common appearance of web search forms. Here, the user types in the keywords to search for as well as a weighting for the different IQ criteria and a budget for the search (Figure 5). With the help of these values the mediator then decides, which and how many search engines to use.

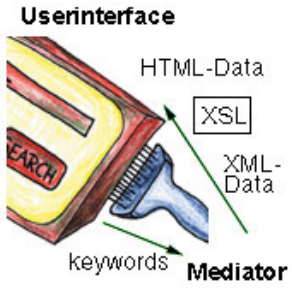

Figure 4: User interface

By the values in the weighting vector the user specifies the individual importance of the different IQ criteria for search engines. The use of a budget is motivated by the idea that the more engines are used for the search, the more time is spent on this process and therefore costs like telephone costs or internet traffic rise. Because search engines do not demand money for their service no other costs occur, but our approach is well applicable to other meta information sources where a charge per query is more common.

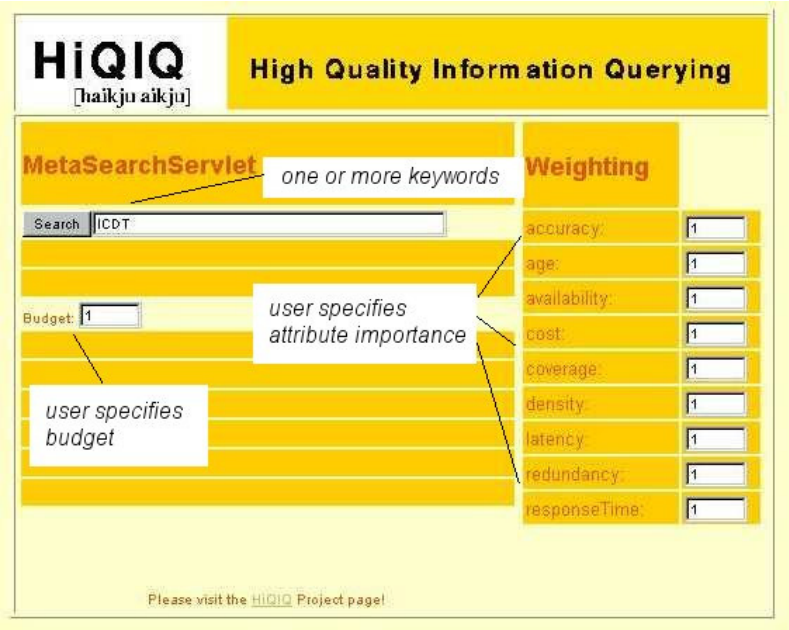


Figure 5: The HiQIQ user interface

**Search results.** The final result of the meta search, which is generated by the mediator, is placed into an XML document and transferred to user interface. There, this data is transformed into HTML with the help of an XSL stylesheet (Figure 4). The user interface itself is a Java servlet which displays the search form (Figure 5) and if available the result of a particular search (Figure 6).

The search results of our HiQIQ meta search engine are valuable for two reasons: They are longer (more links) and they are wider (more attributes). Even search engines that are considered large cover only about 30% of the WWW [LG99]. Thus, meta searching is well worth while—we observe only little overlap and thus are able to respond with a large number of links. This alone does not better the much lamented information overflow. Only a well-ranked response, i.e., a response where the most relevant pages come first, will satisfy a user. To achieve this we rely on two techniques: First, we interleavingly merge ranks of the search engines, because many search engines already have developed sophisticated ranking methods which we make use of. Second, we use our own quality-based ranking as a second level ranking and for conflict resolution within the attributes.

Our results are wider, i.e., we do not return the smallest common denominator (URL, title, and description) as other meta search engines do, but display all available information, thus raising the quality of the result (Figure 6).

# 4 Conclusion and Outlook

The HiQIQ meta search engine is already up and running. The results convince casual users, as informal tests have shown. Still there is much room for further improvements both of technical and qualitative nature.

On the technical side, we plan to include more search engines, improve overall response time, enable flexible reactions when search engines fail, and allow the user to specify the number of results. On the qualitative side, we plan to enhance the weighting interface by (i) storing user profiles to increase usability and (ii) giving quality feedback, e.g., pointing out which weighting led to what decisions of the search engine.

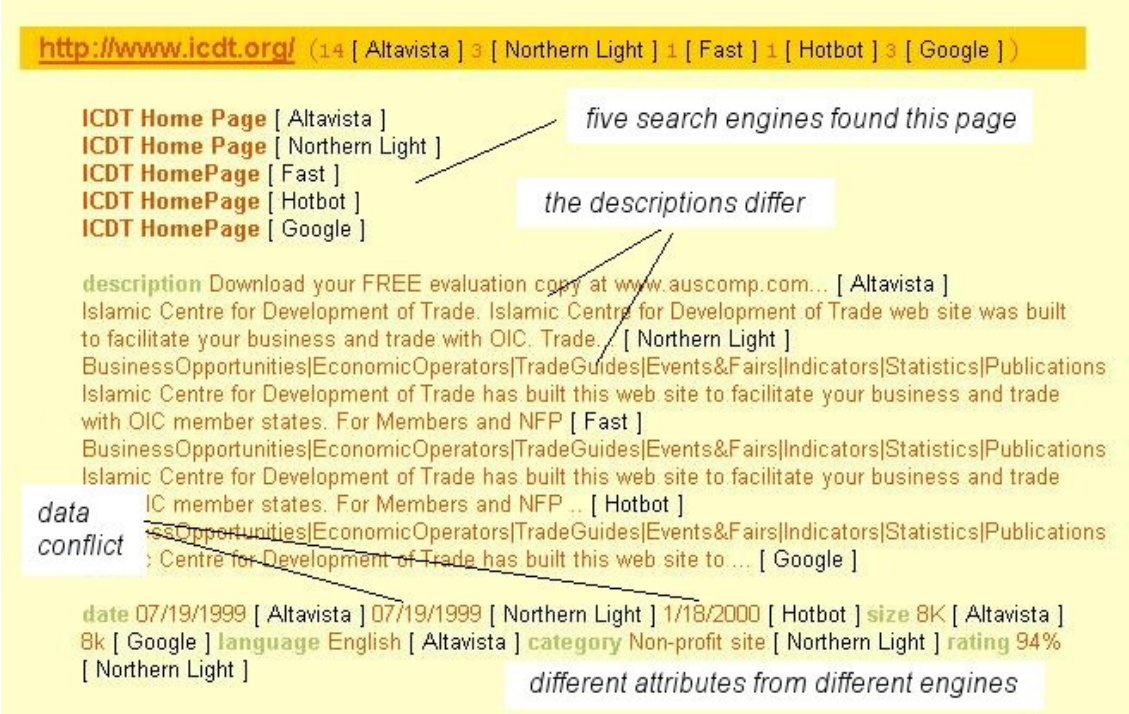


Figure 6: Search result for the keyword "ICDT"

**Acknowledgements.** This research was partly supported by the German Research Society, Berlin-Brandenburg Graduate School in Distributed Information Systems (DFG grant no. GRK 316). We thank Daniel Tonn for the implementation of the ranking methods. For more information on the HiQIQ project please visit `www.hiqiq.de`.

## References

[Coh99] William Cohen. Some practical observations on integration of web information. In *Proceedings of the ACM SIGMOD Workshop on The Web and Databases (WebDB)*, Philadelphia, PA, 1999.

[GGM97] Luis Gravano and Hector Garcia-Molina. Merging ranks from heterogeneous internet sources. In *Proceedings of the International Conference on Very Large Databases (VLDB)*, Athens, Greece, 1997.

[LG99] Steve Lawrence and C. Lee Giles. Accessibility and distribution of information on the Web. *Nature*, 400(6740):107–109, July 1999.

[Nau98] Felix Naumann. Data fusion and data quality. In *Proceedings of the New Techniques & Technologies for Statistics Seminar (NTTS)*, Sorrento, Italy, 1998.

[NLF99] Felix Naumann, Ulf Leser, and Johann Christoph Freytag. Quality-driven integration of heterogenous information systems. In *Proceedings of the International Conference on Very Large Databases (VLDB)*, Edinburgh, 1999.

[NR00] Felix Naumann and Claudia Rolker. Assessment methods for information quality criteria. Technical Report 138, Humboldt-Universität zu Berlin, Institut für Informatik, 2000.

[SA99] Arnaud Sahuguet and Fabien Azavant. Building light-weight wrappers for legacy web datasources using W4F. In *Proceedings of the International Conference on Very Large Databases (VLDB)*, pages 738–741, 1999.

[Wie92] Gio Wiederhold. Mediators in the architecture of future information systems. *IEEE Computer*, 25(3), 1992.