

Techniques for Specialized Search Engines

Robert Steele
Department of Computer Systems
University of Technology, Sydney
PO Box 123 Broadway, NSW 2007 Australia
rsteele@it.uts.edu.au

Abstract

It is emerging that it is very difficult for the major search engines to provide a comprehensive and up-to-date search service of the Web. Even the largest search engines index only a small proportion of static Web pages and do not search the Web's backend databases that are estimated to be 500 times larger than the static Web. The scale of such searching introduces both technical and economic problems. What is more, in many cases users are not able to retrieve the information they desire because of the simple and generic search interface provided by the major search engines.

A necessary response to these search problems is the creation of specialized search engines. These search engines search just for information in a particular topic or category on the Web. Such search engines will have smaller and more manageable indexes and have a powerful domain-specific search interface.

This paper discusses the issues in this area and gives an overview of the techniques for building specialized search engines.

Keywords: specialized search engine, information retrieval, focused crawling, taxonomy, Web search.

1. Introduction

The major search engines such as Google [2], FAST and Alta Vista provide a very valuable resource for searching the Web. However they are not always able to provide the interface to thoroughly search a specialized topic. Also the vast size of the Web with possibly 2 billion Web pages and 550 billion pages in the Invisible Web (Web connected backend databases) [3] is stopping the major

search engines from providing anywhere near complete search.

Even though generic search engines [13] are the most important way for users to currently find information on the Web, there are compelling reasons for the continuing development of specialized search engines. These reasons include both technical/economic advantages and improvements to search ability.

The technical hurdle is that it is becoming increasingly difficult if not impossible for one search engine to index the entire contents of the Web. This is also an economic hurdle – it is not cost-effective given the revenue a search engine can receive to build such an index.

The second reason is the quality of search provided by specialized search engines. The two ways in which specialized search engines can add more to the searching experience are to (1) allow searching of pages that are currently not searchable from the major search engines at all and (2) provide more functionality and search power in the searching of already searchable pages.

The specialized search engines may allow users to search for information that they currently cannot easily search for in a number of ways that will be discussed in Section 2. Broadly speaking this may involve more thorough searching of the static Web or the searching of parts of the Invisible Web (something not currently done by the major search engines). Each specialized search engine can allow searching of just a subset of these pages currently not searchable from the major search engines.

The extra functionality refers to the presenting of interfaces to users that are configured for the particular domain of search. Also the fact that pages of only a targeted category or subject matter appear in the results list can make the user's search experience of a higher quality.

A disadvantage is that people will tend to remember and use just a site that they perceive as being thorough – a search of what might be thought of as a subset of the total information could well be perceived as a waste of time. People don't want to have to go to multiple sites to carry out their searching.

However if a specialized search site is useful enough it will be remembered and used – for example Napster, the music sharing site. Section 4 deals with issues related to integrating the specialized search engines together.

2. Taxonomy

There are two broad ways in which to make a search engine specialized. The first is to build or make use of an index that is itself focused on some particular topic or category of information. The second is to present a search interface to users that supports and implements specialized searches within some category (see Figure 1). In this paper page topic is used to mean subject eg. computing, swimming etc., and category refers to pages of a particular type eg. homepages, FAQs etc. or pages identified by some other cross-topic page characteristic.

These approaches can be used together or independently. If neither is used then the result is a generic search engine such as the current major search engines (see Figure 2).

Specialization of the search index can occur in a number of ways. One way is to retrieve and index only documents that are related to the specific topic or category of interest. Retrieving pages from known relevant sites, by making use of focused crawlers or by using the major search engines to find relevant documents, can do this. A second approach is not to build an index but

to metasearch specialized databases that are relevant to the topic or category of interest. A third way is to do some targeted crawling at query time.

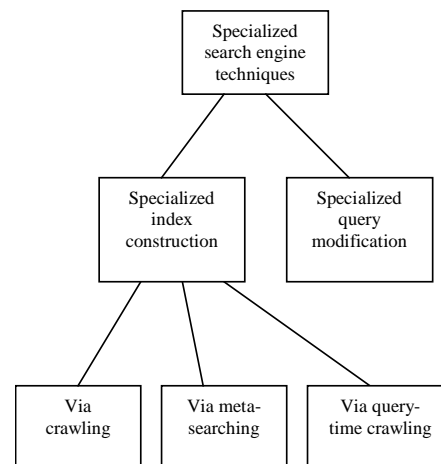


Figure 1. Taxonomy of specialized search engine techniques

Specialization via the search interface and operations on the queries is the other main method for specialization. This can be done in a number of ways. One way is to support particular categories of search by augmenting the user's queries via query modifications that are beneficial for that particular category of search. These searches can make use of specialized indexes or metasearch one or a number of the major search engines. Different specialized search engines can support different specialized searches. Alternatively one search engine can support many categories of search and initiating a specialized search can require the user to explicitly choose that category of search or the category could be inferred from their query.

The rest of Section 2 will present a taxonomy of the different types of specialized search engine techniques.

Subsections 2.1, 2.2 and 2.3 will deal with ways of specialization via the building or use of specialized search engine indexes. Section 2.4 will address specialized query modification.

2.1. Building a specialized index

One way to specialize is to constrain the contents of the search index to a particular topic or category. This section deals with how to build such an index via crawling.

Specialized query modification	[1, 5, 10]	No current systems
No specialized query modification	Generic/ major search engines: Google, Alta Vista	Index building: Flipdog, CiteSeer Meta-searching: Mysimon Query-time
	Generic index	Spedcialized index

Figure 2: Specialized search engine example systems.

Examples of specialized search engines constructed in this way are CiteSeer [12] for the searching of Computer Science papers, Deadliner [8] for finding conference deadlines, HPsearch (<http://hpsearch.uni-trier.de/hp>) for finding personal homepages, Newstracker (<http://nt.excite.com>) and Moreover (<http://www.moreover.com>) that search for the latest news and FlipDog (<http://www.flipdog.com>). In the case of Deadliner the document retrieval phase starts by crawling sites that are known to contain conference deadline postings and retrieving pages from there. Secondly the system makes use of a focused crawler [4, 7] to find many relevant pages. Thirdly the major search engines are metasearched to also find relevant pages. Support Vector Machines (SVMs) can then be trained and then used to more rigorously filter out irrelevant documents from the set to be indexed.

Another search engine that builds a specialized index is FlipDog – it looks at job postings at various IT company Web sites, and builds an up-to-date and powerfully searchable index of job advertisements. FlipDog makes use of the fact that job

advertisements are semi-structured to present the search results in a standard format and allow structured searching. For instance you are able to get a listing of jobs with such constraints as which state the job is in, what category of job it is and what employer is offering that job. No major search engine spiders all of these relevant job pages (certainly not in an up-to-date fashion) or allows such exact specifying of job descriptions. FlipDog does not require manual posting of the job advertisements on their page but rather automatically retrieves and extracts the information.

This demonstrates an important advantage of specialized search engines that have specialized indexes: the content is more structured as it has already been intelligently extracted from Web pages and this allows greater search functionality and information extraction from these pages. For instance Deadliner uses a process called Bayesian Detector Fusion to extract information from the right parts of the documents.

Ways in which crawling by specialized search engines can provide better coverage of the Web than is offered by the major search engines are (1) exploring relevant sites more deeply and (2) providing a more up-to-date search

The major search engines often do not thoroughly search through a Web site that they know of. They may index all pages that are a few links down from the site's home page but may often not go beyond this. This deeper content is then not searchable from the major search engines. A specialized search engine can index the whole of relevant sites.

Another failing of the major search engines is their failure to maintain up-to-date indexes – this is a result of the vast number of pages that must be crawled. A specialized search engine as it has a smaller number of pages to index can more frequently crawl these. The up-to-dateness of the resultant search engine would make it attractive particularly in subject areas where it is important to know of the latest information, for example news. The Moreover (<http://www.moreover.com>) news search

service crawls a set of over 1800 respected news sites some up to four times per hour.

A number of techniques can also be employed to allow such search engines to maintain up-to-date indexes. One of these is for the search engine to monitor the latest postings in relevant newsgroups or relevant bulletin boards. When these postings refer to particular relevant Web sites or to particular topics that are currently discussed at a Web site, these identified Web sites can be re-crawled to check if there have been any changes to the pages at the site.

2.2. Metasearching specialized databases

This is the easiest way to extend the search coverage of search engines [16] but also a second way to specialize the index to be searched. This is another way to allow the searching of content that is not searched by the major search engines.

CompletePlanet [3] estimates the number of searchable databases on the Web to be approximately 200,000. This is too large a number to allow for complete metasearching from one search engine. However it is possible to metasearch a subset of these searchable databases.

These databases can contain very valuable information that is not available elsewhere. For instance groups.google.com (formerly DejaNews) is a searchable database of newsgroup postings. These databases can either contain content found nowhere else on the Web or can contain information pulled together and extracted from other Web pages as is the case with the specialized indexes discussed in Section 2.1.

Metasearching a subset of the databases that are relevant to a particular topic is feasible. This is the approach taken by Mysimon (<http://www.mysimon.com>), the comparison shopping engine. Once the user has narrowed their product category sufficiently a metasearch is made of the vendor sites relevant to that product category and the results incrementally displayed as they are received.

A downside to building a metasearch engine is that the interfaces to the various sites to be searched can change frequently. This means that the metasearcher needs to be continually updated to reflect these changes.

In addition when metasearching a number of related Web databases improving functionality becomes more difficult. One reason for this is that typically the interface to a metasearcher has to take a lowest common denominator approach – it's interface should only allow forms of search that are present in all the search engines queried by the metasearch engine. Also the fact that the different databases will have somewhat different formats means that information extraction becomes difficult.

2.3. Extra crawling at query time

Specialized content can also be found at query-time. This is difficult to do for a generic search engine. There are too many possible sites and pages to attempt to crawl at query-time if the indexed pages are considered inadequate. It may however be possible when the search engine is dealing with a specific topic and has domain-specific knowledge about where to look.

AHOY! [15] is an example of this. If it has not indexed a homepage that you are looking for then using heuristics it may be able to guess the address of the desired home page.

2.4. Specialized search at query time

The three preceding sections have addressed different ways to build or access specialized indexes. This section will address how to make a specialized search engine via the domain specific modifying of queries [11] and query processing. This is also related to what the user understands the search interface to be.

An advantage when search engines search for a particular category of information is that the fact that the domain is constrained leads to word sense disambiguation and can allow a richer search interface functionality.

The type of specialized search engine discussed here does not restrict its searches to a subset of the Web pages available but rather restricts the types of search that can be done. For example Glover et al. [6] used learning techniques to automatically determine successful query modifications to find Web pages in a number of categories such as personal homepages, calls for papers, product reviews and guide or FAQ documents. They used Support Vector Machines to do text classification. This involves training the SVM with positive and negative example pages for the category of interest. From this, query modifications can be learnt so that when a search is made in the future these extra query terms are automatically added to the query. The resultant search engine metasearches generic search engines such as Google and Alta Vista using the original user queries with the extra query terms. As the effectiveness of query modifications are search engine specific, search engine, query modification tuples need to be considered.

Similarly specialized search engines that answer certain types of natural language queries can be developed. This once again uses the learning of query modifications. At query-time the user's query is extended using the learnt query modifications and the modified queries are sent to various major search engines and the results merged.

Agichtein et al. [1] investigate the automatic learning of query modifications for common questions such as "who was...", "where is..." etc. The general idea is to transform questions into specific phrases that may be found on Web pages containing the answer.

These specialized search engines can easily be constructed as metasearch engines that search either one or a number of the major search engines. Alternatively these specialized search abilities could be built into the major search engines. Either the search engines can automatically recognize when a query should trigger a special search (e.g. by noticing a query of the form "What is ...?") or a menu of specialized searches can be provided.

The Geosearch [5] search engine offers search that is aware of the geographical location of resources and factors into the ranking of results whether or not the searcher is in the geographical scope of the resource. This is specialized search where the restricted search category is pages with sufficient geographical relevance. This search is implemented by a second step in query processing that factors in the geographical scope of relevant resources. A lookup table of the geographical scope of all resources is pre-calculated via various algorithms.

3. An objectivity search engine

An interesting category for specialized search that can be dealt with in future work is the searching for objective product review pages. With the large amount of commercial sites presenting biased accounts of products it can be hard to acquire objective information or comparisons about a product. For instance a search for "Jbuilder" on Google returns a page from the Borland Web site – if you were wishing to decide on whether to use Jbuilder or some other competing IDE this would not be the best place to get unbiased information.

While the ability to search for objective pages of all types is interesting it is also ill-defined. The suggestion here is simply for a search engine that finds objective pages about commercial products and pages about how they compare with competing products.

For instance pages that objectively review two competing products may mention both products but also the words "compared with" or "versus".

A basic heuristic for identifying objective articles about products would be to look for pages that mentioned competing products also and preferably a number of times. Either the user can supply a list of competing products or they can supply just one and then the search engine can automate the process of finding competitors. One way to do this would be for the search engine to search for the products name along with such words as "competitor" or "rival" and then intelligently extract the names of competitors. The process

of determining which of these query modifications work best can be automated as has been done in [6]. Alternatively link analysis can be used as is done by GoogleScout (<http://www.google.com>).

Once the competitors are known articles that mention a number of the competitors and the original product along with possible query modifications such as “versus” or “compared” can be searched for. Once again finding the best query modifications for this purpose can be automated.

4. The integration of specialized search engines into general Web search

A big disadvantage with specialized search engines is that people simply want to use one all-purpose search engine. Ways that specialized search engines can be made known to users are (1) a directory of specialized search engines, (2) a metasearch engine of the specialized search engines or (3) the major search engines can automatically infer the specialized search engine(s) that a query should be directed to.

A metasearch engine of the specialized search engines might be particularly effective. It would be expected that such an engine by making use of the best possible specialized search engine for each search would gain an advantage over the major search engines. The metasearch engine could provide a selectable choice of specialized searches for users to choose from and also attempt to automatically determine when to invoke a specialized search based on the query given.

5. The evolution of the Internet in the light of search difficulties

A future scenario where not all web information is easily searchable by all people is quite possible. If a large number of pages no longer are searchable from the most popular search engines this may have effects on the survival of those pages and sites, as

they might not receive sufficient traffic to justify their maintenance.

It could be argued that if search engines index many pages that have interest to very few people (and this costs the search engine money to do) this might not be economically viable and the search engines would need to cut-back on indexing these low interest pages. This might mean that more pragmatic laws might confront the ideal of universal search.

A counter argument is that the major search engines are competing with each other to establish a reputation as the best search engine. This then requires that they all attempt to search as many pages as possible even if very few people are interested in much of the content. Nevertheless once a search engine establishes itself as the best it would then have some leeway to attempt to make its indexing information the most efficient and streamlined as possible.

Despite pages not being indexed by the major search engines it could be argued that if the people who are interested in that information are still able to find the pages via some alternative method (maybe specialized search engines) then that should still be acceptable. Analogously in the physical world it is not possible to search for every business in the world however many small businesses still exist because enough people know of them and find them useful for them to be sustainable.

A future scenario might be that many pages are only searchable from specialized search engines. People interested in such pages use these search engines. Many pages might not be (as is already the case) searchable from the major search engines. Some pages may be searchable from no search engine.

6. Conclusion

The current major search engines are failing to provide ideal search in a number of ways. They cover a relatively small proportion of the static Web pages, their indexes can be significantly out of date, they do not search they generally do not search the

vast number of pages in the Invisible Web and can fail to provide sophisticated search when the user has a specialized category or topic of search in mind.

Specialized search engines alleviate these problems in a number of ways. They can search more of the Web and in a more up-to-date fashion within their domain. They can provide more search functionality, superior search in their domain versus the major search engines in terms of standard retrieval metrics and provide more structure search results.

Ultimately the future of specialized search engines will be driven by technical and economic imperatives.

References

- [1] E. Agichtein, S. Lawrence, L. Gravano. Learning Search Engine Specific Query Transformations for Question Answering. To appear in *Proceedings of WWW10*, Hong Kong, 2001.
- [2] S. Brin, L. Page. The Anatomy of a Large-Scale Hypertextual Web Search Engine. In *Proceedings of WWW7*, Brisbane, Australia, 1998.
- [3] BrightPlanet LLC. The Deep Web: Surfacing Hidden Value. <http://www.completeplanet.com/Tutorials/DeepWeb/index.asp>, 2000.
- [4] S. Chakrabarti, M. van den Berg, B. Dom. Focused crawling: a new approach to topic-specific Web resource discovery. In *Proceedings of WWW8*, Toronto, 1999.
- [5] J. Ding, L. Gravano, N. Shivakumar. Computing Geographical Scopes of Web Resources. In *Proceedings of the 26th VLDB Conference*, Cairo, Egypt, 2000.
- [6] E. Glover, G. Flake, S. Lawrence, W. Birmingham, A. Kruger, C. Lee Giles, D. Pennock. Improving Category Specific Web Search by Learning Query Modifications. In *Symposium on Applications and the Internet*, SAINT 2001.
- [7] Jon Kleinberg. Authoritative sources in a Hyperlinked Environment. *Proceedings 9th ACM-SIAM Symposium on Discrete Algorithms*, 1998.
- [8] A. Kruger, C. Lee Giles, F. Coetzee, E. Glover, G. Flake, S. Lawrence, C. Omlin. DEADLINER: Building a New Niche Search Engine. *Conference on Information and Knowledge Management*, Washington DC, November 6-11, 2000.
- [9] C. Kwok, O. Etzioni, D. Weld. Scaling Question Answering to the Web. To appear in *Proceedings of WWW10*, Hong Kong, 2001.
- [10] S. Lawrence. Context in Web Search. *IEEE Data Engineering Bulletin*, Volume 23, Number3, pp. 25-32, 2000.
- [11] S. Lawrence, C. Giles. Context and Page Analysis for Improved Web Search. *IEEE Internet Computing*, 2(4): 38-46, 1998.
- [12] S. Lawrence, C. Giles, K. Bollacker. Digital Libraries and Autonomous Citation Indexing. *IEEE Computer*, Volume 32, Number 6, pp. 67-71, 1999.
- [13] M. Mauldin. Method for Searching a Queued and Ranked Constructed Catalog of Files Stored on a Network. US Patent 5,748,954, 1998.
- [14] S. Raghavan, H. Garcia-Molina. Crawling the Hidden Web. *Technical Report 2000-36, Database Group, Computer Science Department, Stanford University*, November 2000.
- [15] J. Shakes, M. Langheinrich, O. Etzioni. Dynamic Reference Sifting: A Case Study in the Homepage Domain. In *proceedings of Sixth International Web Conference, WWW6*, 1997.
- [16] Z. Wu, W. Meng, C. Yu, Z. Li. Towards a Highly-Scalable and Effective Metasearch Engine. *Tenth International Web Conference, WWW10*, Hong Kong, May 1-5, 2001.