Specialized Search in Linguistics and Languages

Zhiping Zheng and Gregor Erbach

Computational Linguistics Department
Saarland University
D-66041 Saarbrücken, Germany
zheng@coli.uni-sb.de, gor@acm.org

Abstract. Seven Tones[1] ([13]) is a search engine specialized in linguistics and languages. Its current database, which is stored on a single machine, contains approximately 240,000 indexed web pages about linguistics and languages. Nevertheless, the search engine is designed for a much larger capacity. It has been used in several other systems and is feasible to be transplanted into a distributed computing environment. The performance of Seven Tones in terms of relevance and web page quality is better than Google in the area of linguistics and languages.

Keywords: intelligent crawling, specialized search engine, information retrieval, automated summarization

1. Introduction

There are thousands of specialized search engines on the Web ([5]). These search engines usually are based on indexed databases that are entirely or partially constructed manually ([2]). Seven Tones uses an intelligent web crawler to roam the Internet and locate web pages about linguistics and languages. It uses a high-speeded indexer to automatically index located web pages. The database uses barrel structure ([3]) and sophisticated hash tables to enhance the speed of indexing and search. As the result, Seven Tones reaches a high speed in a single machine environment, and the speed is ensured even if the size of the database expands largely.

Seven Tones also has an HTML document summarizer that will summarize the content of web pages listed in search result.

2. Specialized vs General

General search engines usually exert great effort to the sizes of their databases. The size of the database is usually regarded as a crown jewel asset of a general search engine. Google, which might be the biggest and most powerful search engine at the current stage, claims that it indexes and searches 2,469,940,685 web pages. Many people seem to be convinced that they can find anything in such enormous databases.

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

Many people only use one or several large general search engines to conduct all kinds of search tasks.

The question is, are users getting the best possible search outcomes from these giant search engines? Low recall and low precision have been persistent problems for information retrieval on the Web mainly due to the ways that information is indexed in the databases of Web search engines. Due to the massiveness of information on the web, human indexing with controlled vocabulary becomes a forbidden task. The indexing of web pages has been mainly accomplished with some automated programs in those indexed web pages' natural languages. Even though natural language indexing and retrieval may be more intuitive to many users than controlled vocabulary indexing, rampant synonymy leads to low recall, and rampant polysemy leads to low precision.

According to Linguist Network, Issue 12-866[2], typing the typical linguistic term "morphology" into general search engines, we can get 255,010 hits in Northern Light; 184,000 hits in Yahoo; and 338,000 hits in Google. However, if you check the returned hits one by one, you will find many of them are not truly relevant to linguistics. Such long lists of returned search results can cause extreme difficulties for information seekers, especially for those information seekers who attempt to retrieve as comprehensive as possible information about targeted topic, instead of one or several accurate hits. According to Rosenfeld and Morville ([9]), when an information seeker is presented with more than ten options in a web page, he or she will very likely feel overwhelmed. Long lists of search results with inaccurate hits can easily lead to users frustration, and failure to extract truly relevant information from the search result.

Here is where a specialized search engine can come onto the stage. As a specialized search engine, due to the limit of the subject that it focuses on, Seven Tones indexed only 240,000 web pages at this time, about 1/10,000 of Google size. For a linguist related query, it will return a much shorter list of results. But all items in the list will be related to linguists. This allows a linguistics researcher to avoid the distraction of large number of irrelevant hits from a large general search engine, in turn, increases the likelihood that the researcher will retrieve more relevant information.

3. Seven Tones and Other Specialized Search Engines

Currently, Seven Tones is the unique search engine specialized in linguistics and languages. There is a search engine[3] on Linguist Network[4] that can search for information on linguistics. But it only searches for Linguist List issues archived on Linguist Network. Thus it should be classified as a site search engine.

Seven Tones is different from many other specialized search engines in several aspects.

Algorithms: Technically, search engines with small scale often use some special and simplified algorithms to implement data structures, to index information, or to conduct search. These algorithms usually will lose usefulness when the scales of the search engines are increased to some points. Thus use of these simplified algorithms

[2] http://linguistlist.org/issues/12/12-866.html

[3] http://linguistlist.org/search.html

[4] http://linguistlist.org/

prevents some specialized search engines from being expanded to large sizes. However, Seven Tones uses algorithms for large-scale search engine, and it is ready to be expanded to a much big size and it can be run on distributed computation environments.

Indexing: Specialized search engines normally returns search results with high precision and high quality, because for many specialized search engines, the web pages in their databases are manually or semi-manually selected. The cost incurred by the process of content selection can largely limit the potential sizes specialized search engines can attain, and currency of their content.

The index files of specialized search engines are often entirely or partially constructed manually, sometimes with help from specialized tools. Thus the qualities of indexed web pages are usually very high. On the other hand, because of the manually work included, it is difficult to update the index regularly. It is very common for a user to find mismatches between search hits and the real web pages, or dead links. Seven Tones uses an intelligent crawler to roam the web and judge if a web page is related to linguistics and languages. The qualities of indexed web pages depend on the quality of the algorithm that does the judgment. This makes it possible to build the entire database totally automatically.

Close vs Open: Specialized search engines are often based on a closed set. The structure of the whole set of indexed web pages are designed in advance. Few changes will be made after the search engine starts operating. This is partially because of the manually selection of web pages. Seven Tones works in a different way. It crawls the web and selects relevant web pages all by itself, so it is running as an open system. New web pages and sites will be added into the indexed files automatically after located by the crawler finds them.

4. Main Components

The main components of this system include: 1) An intelligent web crawler to roam the Internet and locate web pages about linguistics and languages; 2) A high speed indexer to index located web pages. 3) A CGI interface that allows users to search the indexed information; 4) An HTML document summarizer. Some other parts include URL validity checker, stemming algorithm, index data structures etc. Figure 1 presents the structure and working processes of Seven Tone.

Intelligent Web Crawler:

The crawler roams the Internet and locates web pages about Linguist and Languages. A set of term vectors ([10]) is trained from a selected web corpus. The crawler will compare a newly detected web page to the series of term vectors and calculate the distances from the new page to each term vector ([6]). The algorithm uses these distance values to judge the relevance of a web page.

The traditional method for vector distance detection is using TF*IDF and cosine similarity measure formula. It didn’t perform well because it needs a big corpus to support. Because the documents are also very likely domain-specific, IDF based on these documents will be biased and domain dependent. We also thought of calculating

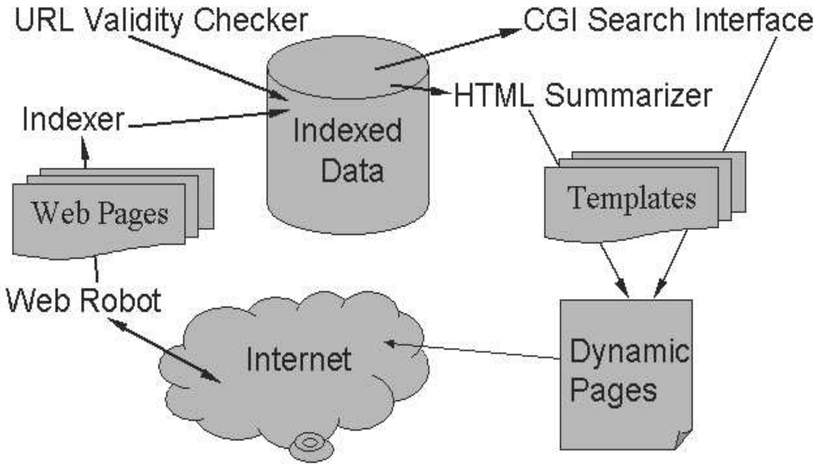

Figure 1. Seven Tones process diagram

the IDF using other widely accepted English corpus like Brown corpus[5], LLC corpus[6] or AP news corpus. But this is even worse because the corpus is domain-independent but the collected web pages are all talking about linguistics and languages.

Currently we are using adjusted term frequency as our main method to form the term vectors. We use square root of frequency as the weight of each term to adjust the term weights so that the importance of a very extensively used term in one document will be balanced. This is our invention and works well. More clearly, we are using the following formula:

$$Sim(X,Y) = \frac{\sum \sqrt{X_i Y_i}}{\sqrt{\sum X_i \sum Y_i}}$$

In this formula, X_i, Y_i are term frequencies in two vectors.

We also uses page importance value, which is similar to PageRank ranking in Google ([7]), to help select qualified web pages.

The web crawler is one of the basic components of this system. The web crawler used in this system is slightly different from web crawlers used for other search engines. It should be smart enough to locate subject-oriented web pages. It not only knows how to judge if a web page is related to a specific topic, but also knows where and how to located those pages efficiently. Currently the crawling algorithm is based on non-recursive level-order tree traversal and also use relevance value to sort the branch walking orders.

High Speed Indexer:

Google search engine uses 64-bit integer addressable virtual files (BigFiles) spinning multiple file system ([3]). This will be easy for the algorithm design but not flexible for the database updates. Most search engines update their databases monthly

[5] http://clwww.essex.ac.uk/w3c/corpus_ling/content/corpora/list/private/brown/brown.html

[6] http://khnt.hit.uib.no/icame/manuals/LONDLUND/INDEX.HTM

or every several months by re-crawling the Web then rebuilding the whole inverted index lists ([4]). Seven Tones uses logically separated files instead of a big virtual file. This makes it possible to update the whole database parallel and part-by-part. As the result, Seven Tones can add a web page to or remove a web page from the index database in minutes.

Only documents related to languages and linguistics will be indexed. The algorithm used in Seven Tones extracts cues from each retrieved web page and compare the cues with a set of calculated and dynamic cues. It assures that most indexed pages in the set are related to the selected topic – linguistics and languages. When you search Seven Tones using term “Shanghai”, you will get “Shanghai Dialect” but will not have much chance to get “Travel to Shanghai.”

Sophisticated hash tables and other data structures are used in the algorithms. The search engine reached high speed in a single machine environment, and the high speed is ensured even if the size of the database expands largely.

Seven Tones indexes web pages mainly in English. A small number of other languages are also accepted. This feature is entailed by a special language recognition function in the indexer.

Search and CGI Interface:

Fuzzy TF-IDF Ranking System: TF-IDF weighting is mostly used in information retrieval research, instead of practical applications. The ranking method used in Seven Tones is derived from normal TF-IDF method.

Seven Tones list 20 URLs on each result page. Search results are sorted by relevance score. Navigation function is put at the bottom. Hyper links to web page itself and its dynamic summarization for each URL.

HTML Document Summarization:

It provides a dynamic summary link for each search result. Other search engines rarely provide a summarization function. Unlike other document summarization algorithms, the algorithm implemented for Seven Tones considers search terms as a factor for summarization. That means the summarizations for the same URL can be different in different searches. Also, the automated summarization is specialized for linguistics and languages.

5. Evaluation

In order to evaluate the performance of Seven Tones, a series of search tasks and assessment of relevance and quality were conducted with Seven Tones and Google. A comparison of the assessments indicates that Seven Tones is superior to Google for search tasks related to linguistics and languages.

To level the search result, traditional method is to compare following three facets of searches: speed, precision, and recall. Modern search engines seem to have solved the speed problem. That means all search engines are fast enough. Thus speed is not considered as a fact affecting search performance here. Recall is also not assessed here for mainly two reasons. 1) When users conduct search tasks, most of them focus on a small number of items that have been listed at the top of the search result queue. Few of them will go through the entire list of returned hits. 2) the amount of workload required to assess the relevance of each returned hit make the task unrealistic to be completed shortly.

While calculating precision, people usually make “yes” or “no” judgment on the relevance of a hit. Nevertheless, web pages can have various extents of relevance to a specific topic. The dichotomy of “yes” and “no” can leave out some useful information for the evaluation of search results.

In addition, the quality of the returned web page is also a piece of information that is relevant to users, given the fact that on the Internet, everyone can be a publisher and, there is little control over the quality of the publications.

Based on the above considerations, this performance evaluation focuses on the relevance and quality of the first 20 hits of a six-point scale to evaluate the extent of the relevance of a piece of search result. Instead of making “yes” or “no” judgment on the relevance of a hit, a six-point scale to evaluate the relevance and quality of the returned web pages was used. The descriptions of the scales are presented in Table 1.

Table 1. Evaluations Scales

Code	Relevance	Quality
5	Well matched	Very Good quality
4	Topic matched	Good
3	Part document matched	Fair
2	Some matched	So-so
1	Word matched	Bad
0	Not matched or dead link	Dead link

When performance score is calculated for each search, each hit’s relevance and quality receives different weights to reflect the fact that users usually pay more attention to the items listed first. The first hit in a search result receive a weight of 2.0, the second hit receive a weight of 1.9, and 20th hit receives a weight of .1. The final relevance score and quality score of a search result are the weighted sums of relevance score and quality of the first 20 hits of the search result.

A list of words is used to compare the search performance of Seven Tones to that of Google. Most words in the list are selected from the keywords that are frequently used by web users in Linguist Network’s search engine. Those words were retrieved from Linguist Network’s server log. Table 2 presents the test results with the words. The mean raw relevance of Seven Tones is 54.4, and the mean raw relevance of Google is 50.4. The mean weighted relevance of Seven Tones is 59.5, and the mean raw relevance of Google is 52.6. The mean raw page quality of Seven Tones is 88.7, and the mean raw page quality of Google is 76.8. The mean weighted page quality of Seven Tones is 90.3, and the mean raw page quality of Google is 80.7. Thus, Seven Tones’ search relevance and quality are both superior to those of Google.

Table 2. Mean Levels of Relevance and Quality of Seven Tones and Google Search Results

	Google	Seven Tones
Raw Relevance	50.4	54.4
Weighted Relevance	52.6	59.5
Raw Page Quality	76.8	88.7
Weighted Page Quality	80.7	90.3

Tables 3 and 4 list a sample of search terms used to conduct the evaluation, and the assessments of the relevance and quality of Seven Tones and Google's search results. From Table 3 and 4, for very specialized linguistics search terms, such as "lingala," "onomastics," "syntax," Seven Tones and Google's search results have similar relevance and quality. Nevertheless, for general terms like "shanghai," "Russian," Seven Tones' search results have much high relevance and quality.

Table 3. Sample Assessment of the Relevance of Google and Seven Tones' Search Results

No.	Term	Raw Relevance		Weighted Relevance	
		Google	Seven Tones	Google	Seven Tones
1	"lingala"	48	49	51.3	52.6
2	"Russian"	26	66	28.9	76.1
3	"shanghai"	18	32	19.1	41.2
4	"syntax"	81	81	81.5	88.9
5	"onomastics"	86	80	88.4	81.3
	Average	51.8	61.6	53.8	68.0

Table 4. Sample Assessment of the Quality of Google and Seven Tones' Search Results

		Raw Page Quality		Weighted Page Quality	
No.	Term	Google	Seven Tones	Google	Seven Tones
1	"lingala"	48	89	51.1	87.3
2	"russian"	92	98	93.4	103.3
3	"shanghai"	66	86	70.1	88.5
4	"syntax"	90	90	93.3	95
5	"onomastics"	87	91	88.2	91.7
	Average	76.6	90.8	79.22	93.16

6. Conclusion and Future Work

Seven Tones is a specialized search engine running on a single UNIX machine. The performance evaluation indicates Seven Tones is superior to Google in retrieving linguistics and languages related information. Nevertheless, this performance evaluation is of a very small scale. The reliability of the relevance and quality assessments is not evaluated. Seven Tones search engine is also used in the local archive version ([14]) of AnswerBus Question Answering System[7] ([11,12]).

Some work from other researchers could be useful for the future work of Seven Tones. For example, [1] and [8] discussed some theoretical and practical approaches for similar classification tasks. Index of word combination (collocation) could be also helpful for the search.

Acknowledgement. The development of Seven Tones was generously supported by Linguist Network.. The author of this paper would like to thank Dr. Anthony Aristar and Dr. Helen Dry for their precious help.

References

[1] M. Alexandrov, A. Gelbukh and P. Makagonov. On Metrics for Keyword-Based Document Selection and Classification. *International Conference on Intelligent Text Processing and Computational Linguistics (CICLing 2000).* Mexico City. February 13-19, 2000.

[2] A. Beavers. Evaluation Search Engine Models for Scholarly Purposes. *D-Lib Magazine.* December 1998.

[3] S. Brin and L. Page. The Anatomy of a Large-Scale Hypertextual Web Search Engine. In *Seventh International World Wide Web Conference.* Brisbane, Australia. April 14-18, 1998.

[4] L. Huang. A Survey On Web Information Retrieval Technologies. Working Paper, http://www.ecsl.cs.sunysb.edu/tr/rpe8.ps.Z. 2000.

[5] S. Lawrence. Context in Web Search. *IEEE Data Engineering Bulletin*, Volume 23, Number 3, pp25-32. 2000.

[6] L. Lee. Measures of Distributional Similarity. In *37th Annual Meeting of the Association for Computational Linguistics, Proceeding of the Conference.* Maryland. June 1999.

[7] L.Page, S. Brin, R. Motwani and T. Winograd. The PageRank Citation Ranking: Bringing Order to the Web. In *Technical Report.* 1998.

[8] R. Prieto-Daz. Implementing faceted classification for software reuse. *Communications of the ACM.* 34(5):89--97, May 1991.

[9] L. B. Resenfeld and P. Morwille. Chapter 3: Organizing information. *In Information Architecture for the World Wide Web.* Cambridge: O'Reilly: 22-46.

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

[10] R. Stata, K. Bharat and F. Maghoul. The Term Vector Database: fast access to indexing terms for Web pages. In *9th International World Wide Web Conference*. Amsterdam. May 15 - 19, 2000.

[11] Z. Zheng. AnswerBus Question Answering System. *Human Language Technology Conference (HLT 2002)*. San Diego, CA. March 24-27, 2002.

[12] Z. Zheng. Developing a Web-based Question Answering System. *The Eleventh World Wide Web Conference (WWW 2002)*. Honolulu, HI. May 7-11, 2002.

[13] Z. Zheng. Seven Tones: Search for Linguistics and Languages. *The 2nd Meeting of the North American Chapter of Association for Computational Linguistics (NAACL 2001)*. Pittsburgh, PA. June 2-7, 2001.

[14] Z. Zheng, H. Huang and S. Schmeier. Deploying Web-based Question Answering System to Local Archive. Fifth International Conference on TEXT, SPEECH and DIALOGUE (TSD 2002). Brno, Czech Republic. September 9-12, 2002.