

Seven Tones

Search in Linguistics and Languages

Zhiping Zheng
School of Information
University of Michigan
zzheng@umich.edu

Abstract

Seven Tones (<http://linguistlist.org/~zheng/7tones/>) is a specialized search engine that allows users to search specifically for information about Linguistics and Languages. It can be easily modified to search information in other topics.

There are thousands of specialized search engines on the Web [4]. These search engines usually are based on indexed databases that are entirely or partially constructed manually [1]. Seven Tones uses an intelligent web crawler to roam the Internet and locate web pages about Linguistics and Languages; and it uses a high-speeded indexer to automatically index located web pages. The database uses barrel structure [2] 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 can be ensured even if the size of the database expands largely. The processes in Seven Tones are presented in Figure 1.

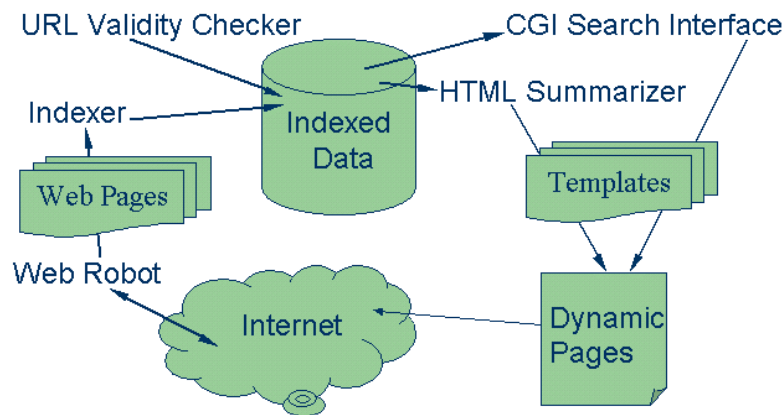


Figure 1. Processes in Seven Tones

Seven Tones has the following key features.

- **Intelligent web crawler:** The crawler roams the Internet and locates web pages about Linguist and Languages. A set of term vectors [7] 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 a new page to each term vector [5]. The algorithm uses these distance values to judge the relevance of a web page. It also uses page importance value (similar to PageRank ranking in Google [6]) to help select qualified web pages.

- **Data structures for indexed files:** Google search engine uses 64-bit integer addressable virtual files (BigFiles) spinning multiple file system [2]. This will be easy for the algorithm design but not flexible for the database update. Most search engines update their database monthly by re-crawling the Web then rebuilding the whole inverted index lists [3]. 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 update its entire database every 15 minutes.
- **Speed:** Sophisticated hash tables and other data structures are used in the algorithms. The search engine reached a high speed in a single machine environment, and the high speed can be ensured even if the size of the database expands largely.
- **HTML Document Summarization:** It provides a dynamic summary link for each search result. Other search engines rarely provide summarization functions. Unlike other document summarization algorithms, the algorithm implemented for Seven Tones considers search terms as a factor for summarization. That means, for different searchers, the summarizations for the same URL can be different.

References

- [1] A. Beavers. Evaluating Search Engine Models for Scholarly Purposes. *D-Lib Magazine*, December 1998
- [2] S. Brin, 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
- [3] L. Huang. A Survey On Web Information Retrieval Technologies. Working Paper, <http://www.ecsl.cs.sunysb.edu/tr/rpe8.ps.Z>
- [4] S. Lawrence. Context in Web Search. *IEEE Data Engineering Bulletin*, Volume 23, Number 3, pp25-32, 2000
- [5] L. Lee. Measures of Distributional Similarity. In *37th Annual Meeting of the Association for Computational Linguistics, Proceeding of the Conference*
- [6] L. Page, S. Brin, R. Motwani, T. Winograd. The PageRank Citation Ranking: Bringing Order to the Web. In *Technical Report*, 1998
- [7] R. Stata, K. Bharat, 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

Acknowledgement

The development of Seven Tones was generously supported by Linguist Network, who provided server space and precious CPU time.