Faster text search with hybrid indexing

Eric Auer - The Language Archive
Max Planck Institute for Psycholinguistics
Nijmegen, The Netherlands

Trova and CQL Search, powered by PostgreSQL and Lucene

The Trova and CQL Search services at The Language Archive allow fast searching for exact (sub-)strings and (at lower speed) even regular expressions. Queries have to return full result lists and statistics on demand, while still showing the first page of results quickly. An elaborate combination of indexes is used to achieve sufficient search performance even for large corpora.

Our current search engine uses a PostgreSQL database and Lucene index in parallel: The database contains all annotation text and file and tier metadata. Tier metadata includes hashed fingerprints of the text content and information about access rights and the DAG tree-like structure in which annotation files are arranged to corpora.

The Lucene index contains all possible substrings of text of each tier up to a given length (e.g. 5-grams) and some tier metadata. There is no position information: Tiers containing “eleph” “lepha” “ephan” and “phant“ may or may not contain “elephant“. This index helps to quickly find a set with ALL tiers where elephants can be found, including some (but not many) which do not actually contain that search term.

Together, both index providers allow to narrow down the set of candidate tiers for a given query a lot. Only those are then scanned in full text, with support for complex queries, regular expressions etc. The search result is not biased by indexing normalizations, index-unfriendly queries are only slower.

Benchmark example: One small server and 110,000,000 strings

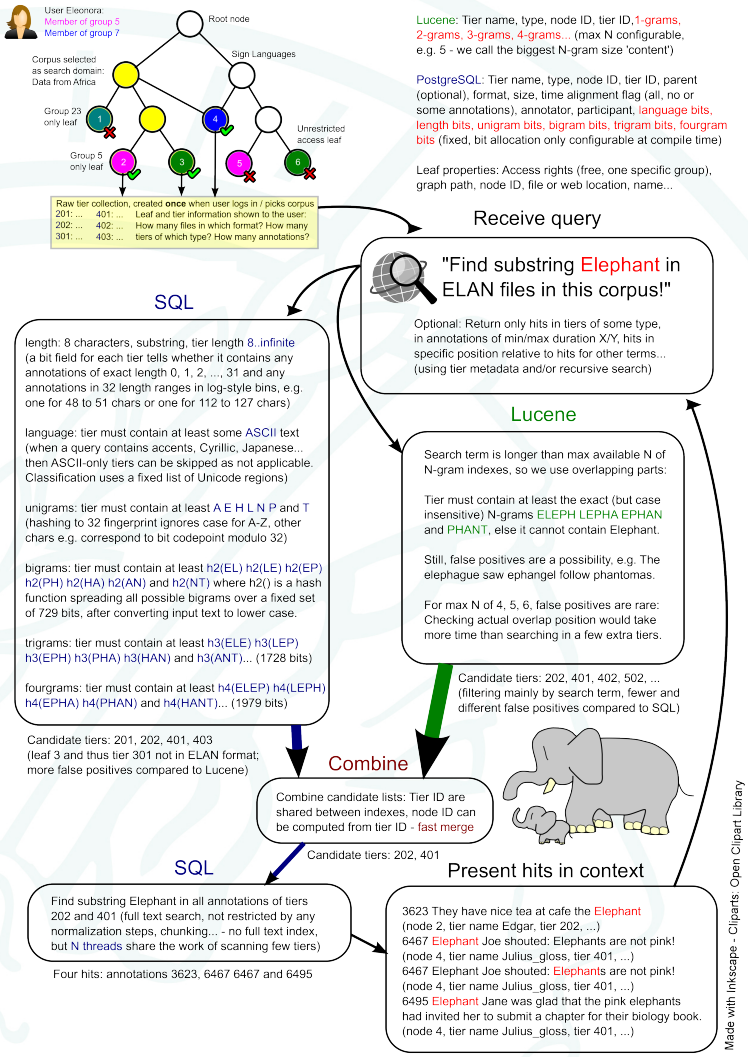
The search engine was tested on an older quad core (2 * Opteron 275 at 2.2 GHz) server with 12 GB of RAM (6 * 2 GB PC2700 ECC) and a small SCSI RAID10 (4 harddisks 10,000 rpm). Size of the test corpus was 100k files with 110M annotations in 700k tiers, a total of 1.8G chars.

At the start of the search session, it takes less than 10 seconds to gather all file metadata from a PostgreSQL DB and Lucene index. Creating those by parsing all annotation files and reading metadata from another database takes less than 3 hours on the same machine. Most frequently used file formats are ELAN EAF, CHILDES CHAT and flat text. The engine also supports Shoe- or Toolbox, HTML, XML, CSV, SubRip SRT, Praat TextGrid, PDF.

Compared to older engine versions, session setup is a few seconds faster now. Search can be focused on (one or more, also checking access rights) smaller subcorpora, which is of course a lot faster for both setup and search. The test corpus is comparable to several bigger real corpora or one “small” archive.

We ran queries for 10 medium frequency (991st - 1000th most frequent) words of 7 languages (D, EN, NL, FR, RU, ES, TR) as well as 20 common Japanese words and 10 manually chosen keywords, e.g. elephants in different languages. The average time to return ALL hits was less than 10 seconds (with 2 or 3 threads, 12 seconds single-threaded) compared to circa 28 seconds without Lucene support. A first page of results was even available after on average only 1.4 seconds, already providing on average 90+ hits. Without Lucene, this task took circa 5 seconds.

Overview of query processing



References

Stehouwer, H., & Auer, E. (2011). **Unlocking language archives using search**. In C. Vertan, M. Slavcheva, P. Osenova, & S. Piperidis (Eds.), Proceedings of the Workshop on Language Technologies for Digital Humanities and Cultural Heritage, Hissar, Bulgaria, 16 September 2011 (pp. 19-26). Shoumen, Bulgaria: Incoma Ltd. http://pubman.mpdl.mpg.de/pubman/item/escidoc:1217575:7

Stehouwer, H., Durco, M., Auer, E., & Broeder, D. (2012). **Federated search: Towards a common search infrastructure**. In N. Calzolari (Ed.), Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC 2012), Istanbul, May 23rd-25th, 2012 (pp. 3255-3259). European Language Resources Association (ELRA). http://pubman.mpdl.mpg.de/pubman/item/escidoc:1478387:2

Contact: eric.auer@mpi.nl