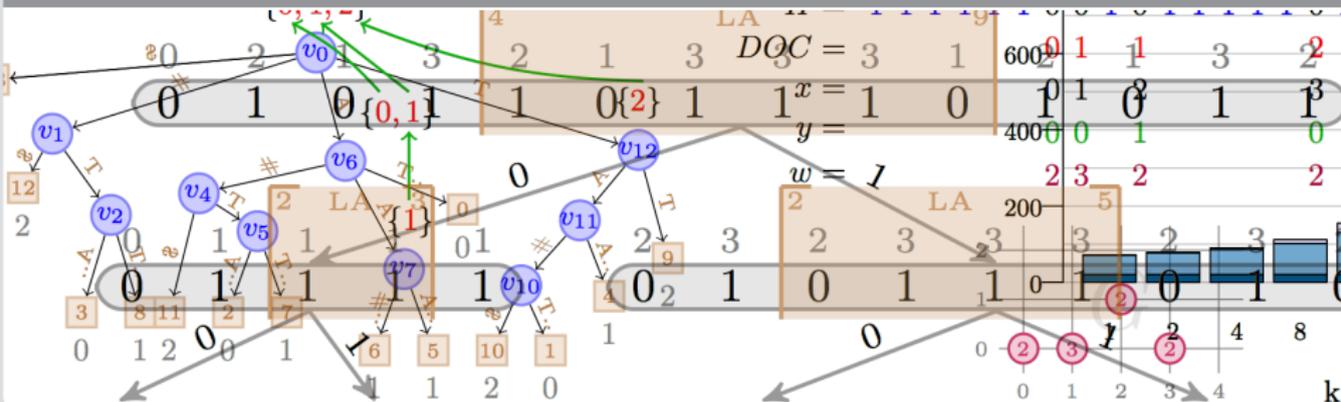


Praxis der Forschung: Efficient Document Retrieval on General Sequences

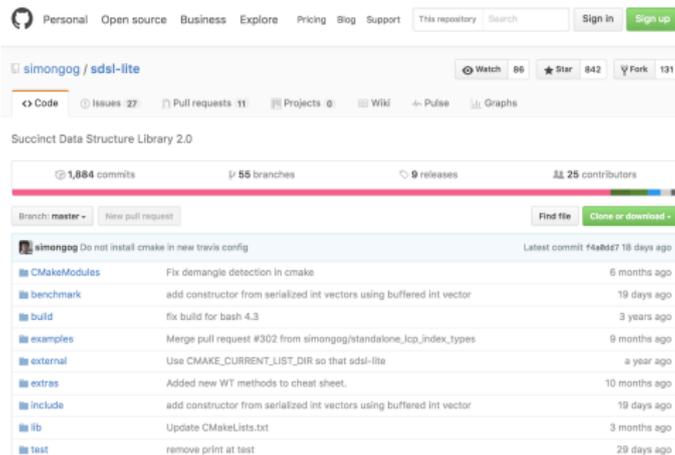
Timo Bingmann, Simon Gog (*gog@kit.edu*), Peter Sanders

Institute of Theoretical Informatics - Algorithmics



- Massive amounts of textual data are publicly available, e.g. WWW data, including source code and DNA databases
- We consider the problem of searching
 - single word or phrase in the data
 - “bag-of-words” queries (like google does)
 - resulting documents are ranked
- Efficient search requires index data structures
- Most search engines (Google, Lucene) are inverted-index based
- We consider a theoretically more attractive solution
 - based on compressed suffix arrays
 - 2d and 3d range search structures

- SDSL (our own library)
 - C++ template library
 - of highly-optimized
 - succinct data structures



The screenshot shows the GitHub repository page for 'simongog/sdsl-lite'. At the top, there are navigation links for 'Personal', 'Open source', 'Business', 'Explore', 'Pricing', 'Blog', and 'Support'. A search bar and 'Sign in'/'Sign up' buttons are also present. The repository name 'simongog / sdsl-lite' is displayed, along with statistics: 96 watchers, 842 stars, and 131 forks. Below this, there are tabs for 'Code', 'Issues 27', 'Pull requests 11', 'Projects 0', 'Wiki', 'Pulse', and 'Graphs'. The repository description is 'Succinct Data Structure Library 2.0'. A progress bar shows 1,884 commits, 55 branches, 9 releases, and 25 contributors. A 'New pull request' button is visible. The file list includes: 'CMakeModules' (6 months ago), 'benchmark' (19 days ago), 'build' (3 years ago), 'examples' (9 months ago), 'external' (a year ago), 'extras' (10 months ago), 'include' (19 days ago), 'lib' (3 months ago), and 'test' (29 days ago).

- Apache Lucene
 - industry standard inverted file based search engine
 - written in Java
 - baseline system



Prerequisites

C++

Java

Algorithms

Are you here?

- Algorithm lectures, text-indexing lecture, advanced data structures, . . .
- Programming skills in C++11 and C++14, Java
- Unix tools (GDB, valgrind, vtune, perf)
- Low level programming
- Scripting languages (bash, python, R)

- Wing-Kai Hon, Rahul Shah, Sharma V. Thankachan, Jeffrey Scott Vitter. *Space-Efficient Frameworks for Top-k String Retrieval*. J. ACM 61(2):article 9, 2014.
- Gonzalo Navarro. Spaces, Trees and Colors: The Algorithmic Landscape of Document Retrieval on Sequences. *Proc. ACM Computing Surveys* 46(4):article 52, 2014.
- Simon Gog and Gonzalo Navarro. Improved Single-Term Top-k Document Retrieval. *Proc. ALENEX*, pages 24-32, 2015.
- Simon Gog, Timo Beller, Alistair Moffat, Matthias Petri. *From Theory to Practice: Plug and Play with Succinct Data Structures*. SEA 2014.