RESEARCH

Efficent Structure Searching with Inverted Indices

Dac-Trung Nguyen*, Rajarshi Guha and Tyler Peryea

*Correspondence:
nguyenda@mail.nih.gov
National Center for Advancing Translational Sciences
National Institutes of Health
9800 Medical Center Drive
Rockville, MD 20850 US
Full list of author information is available at the end of the article

Abstract

The rapid growth of chemical databases has put much burden on our ability to efficiently search and mine chemical structures. Even for modest sized chemical databases (e.g., $\approx 10^6$), supporting full structure searching efficiently remains a challenging task. Herein we describe an efficient indexing scheme—based on inverted indices—for chemical graphs that enables fast structure searching. We demonstrate the utility and effectiveness of our approach through a self-contained implementation available at https://spotlite.nih.gov/opensource/structure-indexer

Introduction

- Short background on chemical structure searching
- Short background on inverted indices
- Prior work on inverted indices - similarity searching [2], structure searching [**?**], clustering [3]
- Contributions from this paper
  - Differentiate from [2]

Methods

- Describe construction of inverted index

Using Lucene as the backend

- Brief background on Lucene
- Why it is it useful to use Lucene rather than homegrown solution?
- integrate structure search along with text/numeric search for free
- Dalke [1] mentions use of Lucene and related libs

Results

Benchmarking query performance

- Query performance can mean execution time and accuracy
- Summarize query performance as function of parameters
  - fingerprint size
  - codebook size
  - other params?

Discussion

- How could other features (distributed queries?) of lucene be leveraged?

Summary

Competing interests
The authors declare that they have no competing interests.

Author's contributions
Text for this section . . .

Acknowledgements
Text for this section . . .

References
1. inverted index 2013.
2. Ramzi Nasr, Rares Vernica, Chen Li, and Pierre Baldi. Speeding up chemical searches using the inverted index: the convergence of chemoinformatics and text search methods. *J. Chem. Inf. Model.*, 52(4):891–900, Apr 2012.
3. Philipp Thiel, Lisa Sach-Peltason, Christian Ottmann, and Oliver Kohlbacher. Blocked inverted indices for exact clustering of large chemical spaces. *J. Chem. Inf. Model.*, 54(9):2395–2401, Sep 2014.

Figures

Figure 1 Sample figure title. A short description of the figure content should go here.

Figure 2 Sample figure title. Figure legend text.

Tables

Table 1 Sample table title. This is where the description of the table should go.

	B1	B2	B3
A1	0.1	0.2	0.3
A2	...	..	.
A3	..	.	.

Additional Files
Additional file 1 — Sample additional file title
Additional file descriptions text (including details of how to view the file, if it is in a non-standard format or the file extension). This might refer to a multi-page table or a figure.

Additional file 2 — Sample additional file title
Additional file descriptions text.