Algorithms for resource-constrained domain-specific knowledge management

Bachelor’s thesis

Uli Köhler

September 24, 2015
Text mining

- Growing number of publications
- Quick access to information required

→ Text mining
Search engines

Google

Search results

Scientific results

Biology & Medicine

Domain of interest
Search engines

Google

PubMed

Search results

Scientific results

Biology & Medicine

Domain of interest
Domain-specific text mining

- Results from outside of area of interest → High false-positive rate
- Large-scale text mining is resource-intensive → Expensive hardware required

- **Domain-specific text mining** covers only a small area of interest
- Smaller dataset → cheap commodity hardware
TRANSLATRON

- **Translational Bioinformatics Tool with realtime ontology**
- A simple tool for domain-specific text mining
- Easy to use — low hardware requirements
- Web-based user interface
- Real-time search in corpus and ontology
- Named entity recognition (NER)

Example: *The PrPSc prion causes ovine prion diseases*
Algorithms for domain-specific text mining

- Conventional algorithms built for large-scale datasets:
  - Hundreds of gigabytes of RAM available
  - Hundreds of terabytes of disk space available
  - Clustered architecture
Algorithms for domain-specific text mining

- Conventional algorithms built for large-scale datasets:
  - Hundreds of gigabytes of RAM available
  - Hundreds of terabytes of disk space available
  - Clustered architecture

- Novel algorithms required for domain-specific approaches
Algorithms in *TRANSLATRON*

- *YakDB* High-performance database
- *PRIMORDIAL* text indexing
- *PRAISER* distributed indexing
- *PERSIST* single-token indexing
- *PRESIDE* real-time prefix search
- *PRO-PANE* priority-based result ordering
- *FiT-NESS* named entity recognition
- *WESTSIDE* client interface
FiT-NESS

- **First-Token-based Named Entity Selection Scheme**
- Trivial: *Single-token entities* like *BRCA1*
- Hard: *Multi-token entities* like *prion diseases*
FiT-NESS

- First-Token-based Named Entity Selection Scheme
- Trivial: *Single-token entities* like *BRCA1*
- Hard: *Multi-token entities* like *prion diseases*
- FiT-NESS approach:
  - Ignore everything but the first token
  - When we find a hit, check if subsequent tokens match the entity
prion diseases | MeSH:D017096

prion  →  prion diseases | MeSH:D017096
"The $PrP^{Sc}$ prion causes ovine prion diseases"
Key advantages of **TRANSLATRON**

- Can be installed on resource-constrained devices:
  - Notebooks
  - Mobile devices (smartphone, tablet, ...)
  - Embedded devices
- Simple architecture
- Easily adaptable to specific requirements
- Can import internal documents (lab reports, ...)
- Individual installations for each researcher or workgroup
Live demonstration
Outlook & conclusion

**TRANSLATRON** is only a proof-of-concept but easier to adapt than conventional tools. Only basic features are implemented and not infinitely scalable. Applications in disaster relief? Applications for text mining with internal documents?
Outlook & conclusion

TRANSLATRON is only a proof-of-concept … but easier to adapt than conventional tools

- Only basic features are implemented
- Not infinitely scalable
- Applications in disaster relief?
- Applications for text mining with internal documents?
Acknowledgements

Mathias C. Walter
Prof. Dr. Hans-Werner Mewes
... and many others ...
Thank you for your attention!

References and sources available at

https://github.com/ulikoehler/Bachelor
https://github.com/ulikoehler/Translatron
https://github.com/ulikoehler/YakDB

Contact: ukoehler@techoverflow.net

Questions?
Image sources

http://www.case.edu/med/nutrition/images/pubmed-logo.jpg
http://mips.helmholtz-muenchen.de/excerbt
http://www.raspberrypi.org/blog/raspberry-pi-2-on-sale/
https://www.raspberrypi.org/blog/raspberry-pi-2-on-sale/
http://www.depts.ttu.edu/hpcc/

Wachinger: Next Generation Knowledge Extraction from Biomedical Literature with Semantic Big Data Approaches
Excerbt architecture
YakDB architecture

Main thread
- Request Router
- Reply Proxy

Client application
- Server Info & ASYNC replies
- REQUEST
- RESP

Log distribution
- Persistent Log Storage

Read worker threads (multiple)
- Read processor
- Shared tablespace
  - RocksDB database (one for each open table)

Update worker threads (multiple)
- Batch processor
- Table processor
- Table operations thread

Legend:
- ZeroMQ Socket
- Fast actor
- Time-intensive actor
- Netlabel

Non-inproc transport
Inproc transport

Request path
Reply path

Log

Table operations thread
Manages mapping from numeric table ID to LevelDB database pointer

Non-inproc transport
Inproc transport

Read-only
Read-write
Open / close table

Lockfree synchronization of table open requests and tablespace resizing

Request path
Reply path

RocksDB database objects automatically synchronize requests

Legend:
- ZeroMQ Socket
- Fast actor
- Time-intensive actor
- Netlabel

(same netlabels connect together)
DISC1 function thus intersects in a complex manner with both of these therapeutically relevant signaling pathways (Figure 4), being modulated through NMDA receptor signaling and in turn affecting the surface expression of both dopaminergic and glutamatergic receptors. Excitingly, DISC1 and its complex may therefore provide an opening for therapeutic modulation of either or both dopaminergic and glutamatergic receptor function and signaling.

DISC1 therapeutic pathways. Neurologically relevant cellular signaling pathways influenced by DISC1 are shown. Green arrows depict activation enzymes, or otherwise enhancement of the target functions (for example, by leading to upregulated transcription of the protein). Red arrows depict inhibition or otherwise downregulation. Black arrows depict effects which do not fall easily into one of the above categories or that are not yet fully understood. Data on the role of DISC1 in these pathways was taken from refs (9, 11, 36, 37, 70, 71, 98, 131, 158, and 200). Dashed arrows indicate indirect effects. Refer Abbreviations for full names and text for further details.