Parsing MetaMap Files in Hadoop

Amy L. Olex, M.S., Alberto Cano, Ph.D., Bridget McInnes, Ph.D.

Department of Computer Science

Hadoop for the Desktop

The **deluge of data** in today's information-centric world requires bigger and better computing resources for processing. This can be a limiting factor in how much data labs with **limited computing resources** are able to handle. This project explores the pitfalls of a serial program that parses MetaMap files to identify UMLS CUI bigrams in a large set of scientific literature. The algorithm is re-implemented in Hadoop MapReduce to **overcome resource bottlenecks**.

UMLS, CUIs, and MetaMap….Oh My!

UMLS[1]

- **Unified Medical Language System**
- Repository of biomedical vocabularies
- Facilitates automated information retrieval *(e.g. linking health information and billing codes across systems)*.
- Provides **language normalization** by linking similar terms to same concept/meaning.

CUI

- **Concept Unique Identifier**
- ID assigned to each **unique concept** in the UMLS.
- E.g. Headache and Cranial Pain are both assigned to the CUI *C0018681*

MetaMap[2]

- Tool that **identifies UMLS concepts** in biomedical texts.
- Output mapped text to compressed MetaMap Machine Output (MMO) files.
- Parsed 23,343,329 citations to create the 2015 MedLine/PubMed Baseline dataset--779 MMO files compressed to 132GB.

UMLS::Association

UMLS CUIs can be used to normalize biomedical and clinical text for use in natural language processing applications. By counting **CUI Bigram frequency**—the number of times two CUIs appear close to each other in text--the UMLS::Association package can **identify related concepts**. E.g. head ache and aspirin.

MetaMap File Anatomy

Utterance
16691646.ti.1 *"Statement of Cases of Gonorrhoeal and Purulent Ophthalmia…"*

Phrase 1
"Statement of Cases of Gonorrhoeal"

Mapping	"Statement"	"Cases"	"Gonorrhoeal"	
	C1710187	C0868928	C0018081	'Case (situation)'
Mapping	"Statement"	"Cases"	"Gonorrhoeal"	
	C1710187	C1533148	C0018081	'Case unit dose'

Phrase 2
"Purulent Ophthalmia"

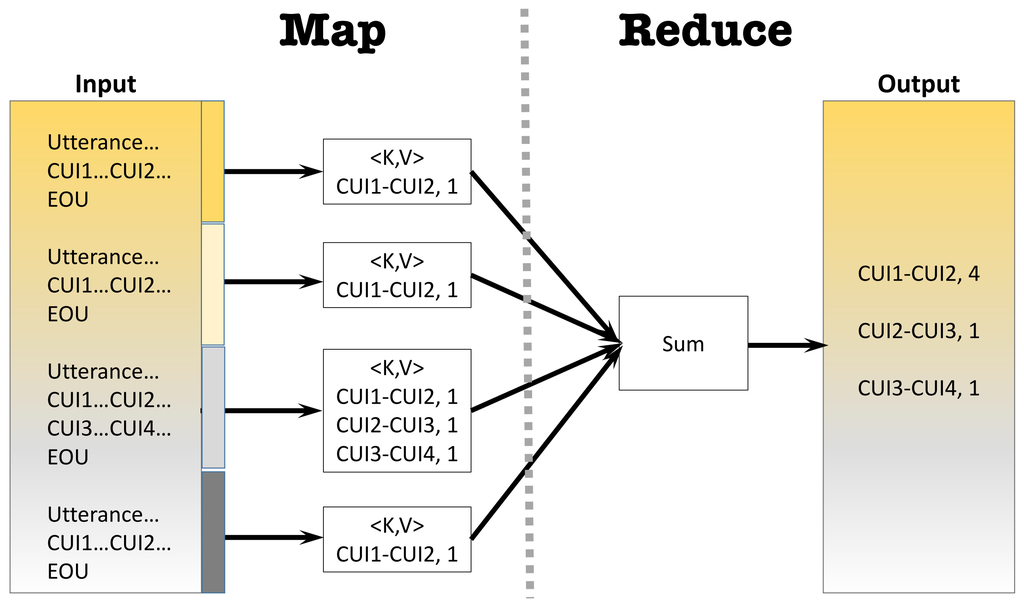
Mapping	"Purulent Ophthalmia"
	C0259800

Adjacent CUI Bigrams
C1710187 - C0868928
C0868928 - C0018081
C0018081 - C0259800
C1710187 - C1533148
C1533148 - C0018081

Serial Limitations

- Processes **one file/one utterance** at a time with nested for loops.
- Regularly writes nested bigram hash table to MySQL database due to **memory limitations**, introducing **DB communication latency**.
- Perl code is **not paralellizable** due to limitations in sharing nested hashes across threads.

MapReduce in 30 seconds



Split input by record --> identify **<Key, Value>** pairs in parallel --> sum values with same key

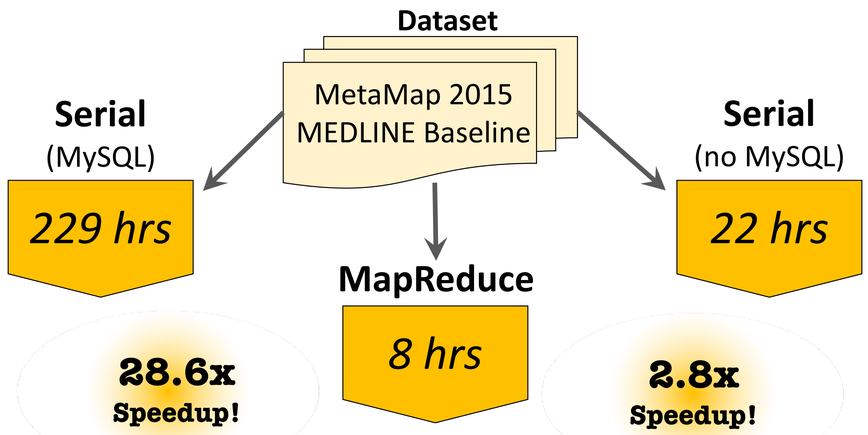
MapReduce Advantages

- Inherent and **scalable parallelization**.
- **Writes results to disk**--all intermediate and final.
- **No MySQL communication**.

Results

CuiCollectorMapReduce

Extracts CUI bigrams using Hadoop MapReduce framework.
Desktop implementation (a single node Hadoop system).



Conclusion and Future Work

Parsing CUI Bigrams in Hadoop on a desktop computer resulted in **significant speedup**, and enables **parsing larger datasets** that were not previously feasible. Algorithm improvements include a window size to collect distant CUI bigrams and crossing utterances to process full PubMed citations. Future work includes testing the scalability on a Hadoop cluster, resolving an issue with the compressed input file format to improve mapping efficiency, identifying optimal Hadoop settings for a desktop implementation, and re-implementing in SPARK to take advantage of its in-memory storage of intermediate results.

In conclusion, **desktop implementations** of Hadoop can **resolve computing resource problems** and **process data faster**, opening up **more research areas** in big data processing for smaller labs.

References

[1] Bodenreider O. "The unified medical language system (UMLS): integrating biomedical terminology." *Nucleic acids research*, 2004, 32, D267-D270.
[2] Aronson AR and Lang F. "An overview of MetaMap: historical perspective and recent advances." *JAMIA*, 2010, 17:3, 229-236.

Contact alolex@vcu.edu or btmcinnes@vcu.edu for more information.