MapReduce in the Cloud: A study for efficient co-occurrence processing of MEDLINE annotations with MeSH

Markus Kreuzthaler, Jose Antonio Miñarro-Giménez and Stefan Schulz

Institute for Medical Informatics, Statistics and Documentation, Medical University of Graz, Austria

HEC 2016, 30.08.2016, MI-2-5 Advanced analytics and Big Data
Introduction

- PubMed/MEDLINE +24M records
- Scopus +55M records
- ScienceDirect +12M records
- BioMed Central
- Medscape
- Google Scholar
- HighWire +7M records
Knowledge Types

Knowledge in biomedical terminologies

- **Lexical**: “cancer, “carcinoma”, “Krebs” have the same meaning.
- **Ontological**: “lung cancer is a cancer”, “lung cancer is located in the lung”, representing what is universally true.

Contingent knowledge

- **Context-dependent**: Sudden fever may be highly indicative for malaria in Sub-Saharan Africa but not in Central Europe.
- **Probabilistic**: Smokers have a higher risk for lung cancer
- **Subject to temporal change**: A drug was indicated to treat a certain disease in the past, but it is used for different purpose today, or has been withdrawn from the market.
MEDLINE

- Citations and abstracts from biomedical literature.
- Contains +26 Million records maintained by NCBI at the U.S. NLM.
- Index terms using the MeSH thesaurus support literature search, optimized regarding precision and recall.

Medical Subject Headings (MeSH)
  - Controlled vocabulary
    - ~27,800 descriptors
    - ~87,000 entry terms
  - Hierarchical structure
  - Continually revised and updated
  - Used for manual indexing of each publication in MEDLINE
Pathophysiology and clinical relevance of Helicobacter pylori.

Halter F¹, Hurlimann S, Inauen W

Abstract
Considerable knowledge has recently accumulated on the mechanism by which Helicobacter pylori (H. pylori) induces chronic gastritis. Although H. pylori is not an invasive bacterium, soluble surface constituents can provoke pepsinogen release from gastric chief cells or trigger local inflammation in the underlying tissue. Urease appears to be one of the prime chemoattractants for recruitment and activation of inflammatory cells. Release of cytokines, such as tumor necrosis factor alpha, interleukin 1 and 6, and oxygen radicals, leads to a further tissue inflammation accompanied by a potent systemic IgA and IgG type of immune response. Chronic inflammation and antigens on glandular epithelial cells lead to a progressive destruction with loss of the epithelial barrier function. Within the gastric mucosa, patches of intestinal metaplasia develop, which may be a risk factor for subsequent development of gastric carcinoma. Hyperacidity in duodenal ulcer patients induces gastric metaplasia in the duodenal bulb, which represents a target for H. pylori colonization and ulcer formation. H. pylori can be detected in the majority of patients with peptic ulcers and, compared to age-matched healthy people, it is also found more often in patients with dyspepsia and gastric carcinoma. Although H. pylori can be detected in healthy people, the marked reduction of the ulcer recurrence rate by eradication of H. pylori (80 percent versus 20 percent relapse within one year) suggests that H. pylori is a major risk factor for duodenal ulcer formation. The potential role of H. pylori in non-ulcer dyspepsia and carcinogenesis is under investigation. Current regimens aimed at eradicating H. pylori use a combination of several drugs that are potentially toxic. Since the risk of complications may exceed the potential benefit in most patients, eradication treatment should be limited to clinical trials and to patients with aggressive ulcer disease. New drug regimens, e.g., the combination of proton pump inhibitors with one antibiotic, may provide less toxic alternatives. Beyond ulcer treatment, effective and well-tolerated eradication regimens may have a place in prophylaxis of gastric carcinoma.

PMID: 1341068 [PubMed - indexed for MEDLINE] PMCID: PMC2589759 Free PMC Article
Pathophysiology and clinical relevance of Helicobacter pylori.

AU - Halter, F
AU - Holter P
DD - Gastrointestinal Unit, University Hospital, Inselspital, Bern, Switzerland

MH - Gastritis/etiology/physiopathology
MH - Gastrointestinal Diseases/*etiology/*physiopathology
MH - Helicobacter Infections/*complications
MH - Helicobacter pylori/*physiology
MH - Humans
MeSH - Subheadings

- 84 MeSH subheading types for refining the meaning of main headings

**AB**  Abnormalities
**AD**  Administration and Dosage
**AE**  Adverse Effects
**DT**  Drug Therapy
**TU**  Therapeutic Use
...

- Can be seen as **sparse feature vector per co-occurrence**.
- The co-occurrence is seen as **a point in the 84 dimensional subheading space**.
Exploiting **co-occurrence** information together with **subheading annotations** provided by **MeSH** an additional knowledge layer can be build constituted by <SUBJ, PRED, OBJ> triples with predicates like:

<table>
<thead>
<tr>
<th>Disease/Syndrome</th>
<th>Finding</th>
<th>Substance</th>
<th>Organism</th>
</tr>
</thead>
<tbody>
<tr>
<td>complicates causes co-occurs with</td>
<td>occurs in diagnoses</td>
<td>treats prevents causes occurs in</td>
<td>affected by causes</td>
</tr>
<tr>
<td>Disease/Syndrome</td>
<td>Finding</td>
<td>Substance</td>
<td>Organism</td>
</tr>
<tr>
<td>produces diagnosed by</td>
<td>complicates causes co-occurs with</td>
<td>treats prevents causes</td>
<td>affects caused by</td>
</tr>
<tr>
<td>Finding</td>
<td>Substance</td>
<td>Organism</td>
<td></td>
</tr>
<tr>
<td>caused by treated by prevented by diagnosed by</td>
<td>treated by caused by prevented by</td>
<td>interacts is affected by produces</td>
<td></td>
</tr>
<tr>
<td>Substance</td>
<td>Organism</td>
<td></td>
<td></td>
</tr>
<tr>
<td>caused by affects</td>
<td>affects produced by</td>
<td>interacts with</td>
<td></td>
</tr>
</tbody>
</table>
Limitations:

- Focus on the semantic types *Disease/Syndrome, Pharmacologic Substance*

<table>
<thead>
<tr>
<th>Disease/Syndrome</th>
<th>Finding</th>
<th>Substance</th>
<th>Organism</th>
</tr>
</thead>
<tbody>
<tr>
<td>complicates causes co-occurs with</td>
<td>occurs in diagnoses</td>
<td>treats prevents causes occurs in</td>
<td>affected by causes</td>
</tr>
</tbody>
</table>

- Limit data set to MEDLINE records published in the last 5 years
- Concepts are flagged as major topic
1. **Aggregate** co-occurring concept pairs and their subheading vectors

UMLS MRCOC table as main processing resource ($>10^9$ entries, 130 GB)
Material and Methods

2. **Calculate** the corresponding log-likelihood ratio scores (LLRs)

<table>
<thead>
<tr>
<th>Co-occurrence</th>
<th>CUI1</th>
<th>¬CUI1</th>
</tr>
</thead>
<tbody>
<tr>
<td>CUI2</td>
<td>#CUI1_CUI2</td>
<td>#¬CUI1_CUI2</td>
</tr>
<tr>
<td>¬CUI2</td>
<td>#CUI1_¬CUI2</td>
<td>#¬CUI1_¬CUI2</td>
</tr>
</tbody>
</table>

*CUI = UMLS concept identifier*

\[
H = -\sum_{i=1}^{n} p_i (\log_b p_i)
\]

\[
\text{LLR} = 2 \ (H(\text{matrix}) - H(\text{rows}) - H(\text{cols}))
\]

- \(H(\text{matrix})\) Matrix entropy
- \(H(\text{rows})\) Sum of row entropies
- \(H(\text{cols})\) Sum of column entropies
Material and Methods

JobFlow

map \( (k_1, v_1) \rightarrow \) list \( (k_2, v_2) \)

reduce \( (k_2, \text{list}(v_2)) \rightarrow \text{list}(v_2) \)

Step 1 → Step 2 → Step 3 → Step 4 → Step 5 → Step 6

amazon S3

MRCOC

> 10^9 entries

~130GB

EC2 Master

EC2 Slaves

amazon web services™ Elastic Map Reduce

Mahout

Hadoop

Medical University of Graz, Auenbruggerplatz 2, A-8036 Graz, www.medunigraz.at
Material and Methods

Initial Filtering and Accumulation (IFAA)

map (k1, v1) \rightarrow list (CUI1_CUI2, SH_i)
reduce (CUI1_CUI2, list (SH_1, SH_1, SH_i ..., SH_n)) \rightarrow list (SH_1, SH_1, SH_i ..., SH_n)
list (SH_1, SH_1, SH_i ..., SH_n) : (#SH_1, #SH_i ..., #SH_n)

<table>
<thead>
<tr>
<th>CUI1</th>
<th>CUI2</th>
<th>Subheadings</th>
<th>#CUI1_CUI2</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>n=84</td>
<td></td>
</tr>
</tbody>
</table>

Co-occurrence

<table>
<thead>
<tr>
<th>Co-occurrence</th>
<th>CUI1</th>
<th>¬CUI1</th>
</tr>
</thead>
<tbody>
<tr>
<td>CUI2</td>
<td>#CUI1_CUI2</td>
<td>#¬CUI1_CUI2</td>
</tr>
<tr>
<td>¬CUI2</td>
<td>#CUI1¬CUI2</td>
<td>#¬CUI1¬CUI2</td>
</tr>
</tbody>
</table>
Intermediate Occurrence Calculations (IMOC)

- Step 2: Overall counts $\text{OC}$
  \[
  \text{map} (k_1, v_1) \rightarrow \text{list} (\text{OC}, 1) \\
  \text{reduce} (\text{OC}, \text{list} (1, 1, \ldots, 1)) \rightarrow \text{list} (1, 1, \ldots, 1) : #1
  \]

- Step 3: $\text{CUI1}$ counts $\#\text{CUI1}$
  \[
  \text{map} (k_1, v_1) \rightarrow \text{list} (\text{CUI1}, 1) \\
  \text{reduce} (\text{CUI1}, \text{list} (1, 1, \ldots, 1)) \rightarrow \text{list} (1, 1, \ldots, 1) : #1
  \]

- Step 4: $\text{CUI2}$ counts $\#\text{CUI2}$
  \[
  \text{map} (k_1, v_1) \rightarrow \text{list} (\text{CUI2}, 1) \\
  \text{reduce} (\text{CUI2}, \text{list} (1, 1, \ldots, 1)) \rightarrow \text{list} (1, 1, \ldots, 1) : #1
  \]
Material and Methods

Intermediate Occurrence Calculations (IMOC)

Step 5: Reduce Side Join on CUI1

\[ \#\text{CUI1} \neg \text{CUI2} = \#\text{CUI1} - \#\text{CUI1}_\text{CUI2} \]

\[
\begin{array}{c|c|c|c|c}
\text{CUI1} & \text{CUI2} & \text{Subheadings} & \#\text{CUI1}_\text{CUI2} & \text{CUI1} \\
n=84 & & & & \text{#CUI1} \\
\end{array}
\]

\[ \#\text{CUI1} \neg \text{CUI2} = \#\text{CUI1} - \#\text{CUI1}_\text{CUI2} \]

\[
\begin{array}{c|c|c|c|c|c|c}
\text{CUI1} & \text{CUI2} & \text{Subheadings} & \#\text{CUI1}_\text{CUI2} & \#\neg\text{CUI1}_\text{CUI2} & \text{CUI1} & \neg\text{CUI1} \\
n=84 & & & & & \text{#CUI1}_\neg\text{CUI2} & \#\neg\text{CUI1}_\neg\text{CUI2} \\
\end{array}
\]

Co-occurrence | CUI1 | \neg\text{CUI1}  \\
---------------|------|----------------|
\text{CUI2}   | \#\text{CUI1}_\text{CUI2} | \#\neg\text{CUI1}_\text{CUI2}  \\
\neg\text{CUI2}| \#\text{CUI1}_\neg\text{CUI2} | \#\neg\text{CUI1}_\neg\text{CUI2}  \\

Material and Methods

Final Log-Likelihood Calculation (FLLC)

- Step 6: Reduce Side Join on CUI2

\[
\neg \text{CUI}_1 \text{CUI}_2 = \#\text{CUI}_2 - \#\text{CUI}_1 \text{CUI}_2
\]

\[
\neg \neg \text{CUI}_1 \neg \text{CUI}_2 = \text{OC} - \#\text{CUI}_1 \text{CUI}_2 - \#\neg \text{CUI}_1 \neg \text{CUI}_2 - \#\neg \text{CUI}_1 \text{CUI}_2
\]
Material and Methods

Final Log-Likelihood Calculation (FLLC)

\[ H = -\sum_{i} p_i \log_b p_i \]

\[ \text{LLR} = 2 (H(\text{matrix}) - H(\text{rows}) - H(\text{cols})) \]

H(\text{matrix}): Matrix entropy; H(\text{rows}): Sum of row entropies; H(\text{cols}): Sum of column entropies

<table>
<thead>
<tr>
<th>Co-occurrence</th>
<th>CUI1</th>
<th>¬CUI1</th>
</tr>
</thead>
<tbody>
<tr>
<td>CUI2</td>
<td>#CUI1_CUI2</td>
<td>#¬CUI1_CUI2</td>
</tr>
<tr>
<td>¬CUI2</td>
<td>#CUI1¬CUI2</td>
<td>#¬CUI1¬CUI2</td>
</tr>
</tbody>
</table>

Medical University of Graz, Auenbruggerplatz 2, A-8036 Graz, www.medunigraz.at
Results

Experimental setup

- Amazon instance information: Name: M1 General Purpose Medium; API Name: m1.medium; Memory: 3.75 GB; Compute Units (ECU): 2 units; Cores: 1 core; Storage: 410 GB; Arch: 32/64 bit.

<table>
<thead>
<tr>
<th>Slave instances</th>
<th>IFAA</th>
<th>IMOC</th>
<th>FLLC</th>
<th>Sum</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>50</td>
<td>27</td>
<td>36</td>
<td>113</td>
</tr>
<tr>
<td>4</td>
<td>29</td>
<td>13</td>
<td>17</td>
<td>59</td>
</tr>
<tr>
<td>10</td>
<td>16</td>
<td>9</td>
<td>7</td>
<td>32</td>
</tr>
</tbody>
</table>

Table 1. Processing time (minutes) depending on the number of instances and calculation step. IFAA = Initial filtering and accumulation; IMOC = Intermediate occurrence calculations; FLLC = Final log-likelihood calculation.

- The task was **not feasible on a single desktop machine** without map/reduce applied.
Conclusion and Outlook

- Big Data approach (Amazon EC2, S3; Hadoop, Apache Mahout)
  - Creation of an additional format of MRCOC which can be used by the scientific community in the future.
  - Virtualization on demand 10$
  - Buying dedicated hardware >>>>> 10$

- Some results
  - Rash *is associated with* Antineoplastic Drugs; LLR=60.2
    - chi-squared test, f=1, p<0.001, LLR>10.83
  - Rash *is caused by* Antineoplastic Drugs (accuracy 0.85)
    - clustering of subheading information [2]

- Process abstracts
  - NLP, SemRep, Subheading information [1]
  - Use of Spark with uimaFIT in the future (DKPro)
References and Acknowledgements

This work was performed as a part of the BMFacts project (BMFacts: Knowledge acquisition for a biomedical fact repository), funded by the Austrian Science Fund (FWF): [M 1729-N15].

