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

Medical University of Graz

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 ling 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A drug was indicated to treat a certain disease in the past, but it is used for different purpose today, or has been withdrawn from the market.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**

MEDLINE - PubMed

NCBI Resources How To

PubMed.gov
US National Library of Medicine
National Institutes of Health

PubMed | Search

Create RSS Create alert Advanced

Abstract | Send to:

Yale J Biol Med. 1992 Nov-Dec;65(6):625-38.

Pathophysiology and clinical relevance of Helicobacter pylori.

Halter F[1], Hurlimann S, Inauen W.

Author information

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**

MEDLINE - PubMed

PMID- 1341068
OWN - NLM
STAT- MEDLINE
DA - 19940106
DCOM- 19940106
LR - 20100907
IS - 0044-0086 (Print)
IS - 0044-0086 (Linking)
VI - 65
IP - 6
DP - 1992 Nov-Dec
TI - Pathophysiology and clinical relevance of Helicobacter pylori.
PG - 625-38
AB - [Content of abstract]
FAU - Halter, F
AU - Halter F
AD - 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

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

RF - 107
PMC - PMC2589759
OID - NLM: PMC2589759
EDAT- 1992/11/01
MHDA- 1992/11/01 00:01
CRDT- 1992/11/01 00:00
PST - ppublish
SO - Yale J Biol Med. 1992 Nov-Dec;65(6):625-38.

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:

	Disease/ Syndrome	**Finding**	**Substance**	**Organism**
Disease/ Syndrome	complicates causes co-occurs with	occurs in diagnoses	treats prevents causes occurs in	affected by causes
Finding	produces diagnosed by	complicates causes co-occurs with	treats prevents causes	affects caused by
Substance	caused by treated by prevented by diagnosed by	treated by caused by prevented by	interacts	is affected by produces
Organism	caused by affected by	caused by affects	affects produced by	interacts with

Medical University of Graz

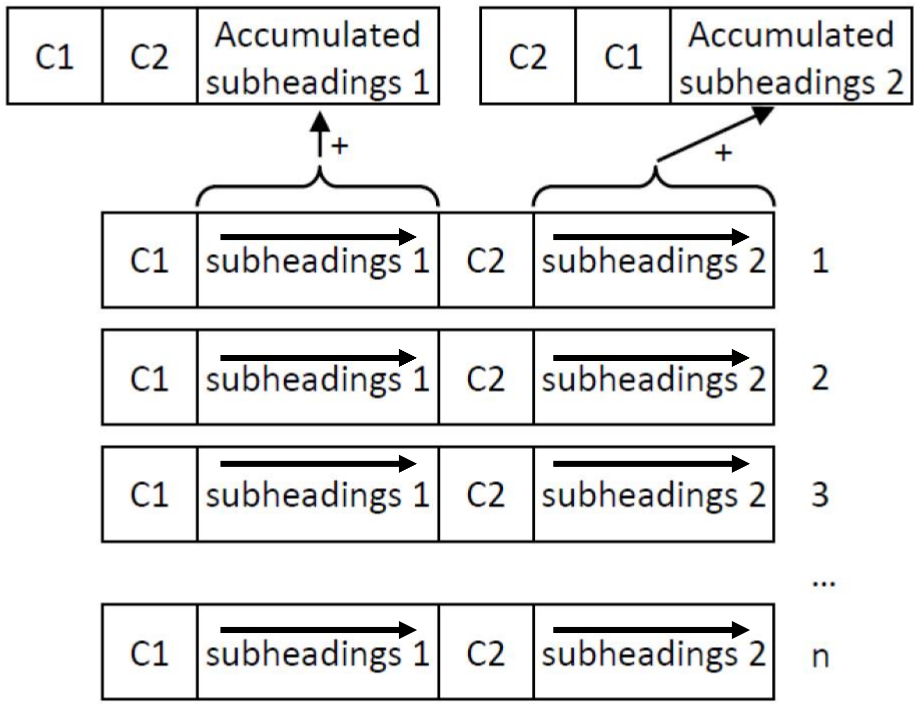
- Limitations:
  - Focus on the semantic types *Disease/Syndrome*, *Pharmacologic Substance*

	Disease/ Syndrome	Finding	Substance	Organism
Disease/ Syndrome	complicates causes co-occurs with	occurs in diagnoses	treats prevents causes occurs in	affected by causes
Finding	produces diagnosed by	complicates causes co-occurs with	treats prevents causes	affects caused by
Substance	caused by treated by prevented by diagnosed by	treated by caused by prevented by	interacts	is affected by produces
Organism	caused by affected by	caused by affects	affects produced by	interacts with

  - Limit data set to MEDLINE records published in the last 5 years
  - Concepts are flagged as major topic

Material and Methods

1. **Aggregate** co-occurring concept pairs and their subheading vectors



UMLS MRCOC table as main processing resource (**>10^9 entries, 130 GB**)

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

Co-occurence	CUI1	¬CUI1
CUI2	#CUI1_CUI2	#¬CUI1_CUI2
¬CUI2	#CUI1_¬CUI2	#¬CUI1_¬CUI2

CUI = UMLS concept identifier

$$H = -\sum_{i}^{n} p_i(log_b p_i)$$

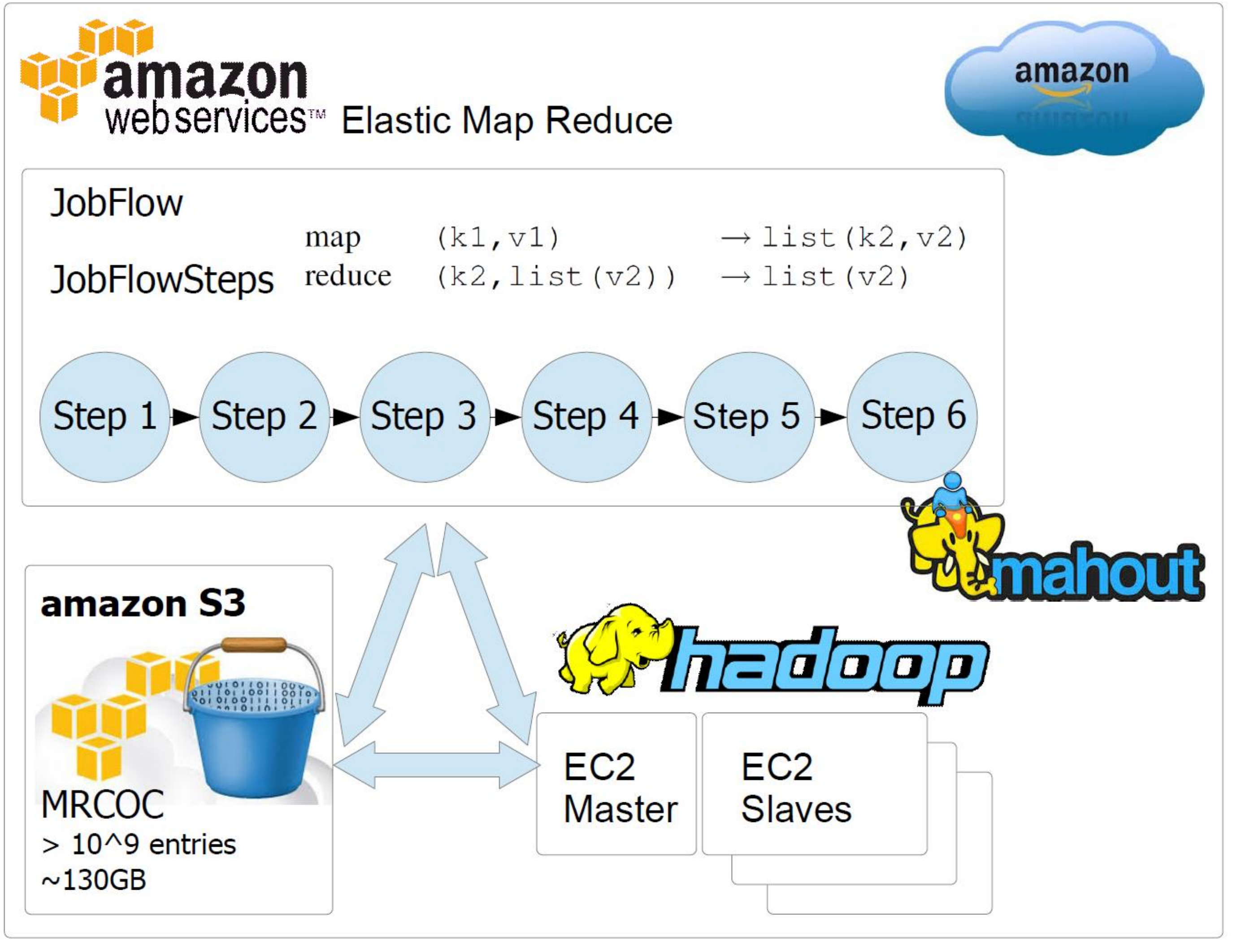
LLR = 2 (H(matrix) - H(rows) - H(cols))

H(matrix) Matrix entropy

H(rows) Sum of row entropies

H(cols) Sum of column entropies

Material and Methods



Material and Methods

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

Initial Filtering and Accumulation (IFAA)

map (k1, v1) → list (CUI1_CUI2, SH_i)

reduce (CUI1_CUI2, list (SH_1, SH_1, SH_i …, SH_n)) → list (SH_1, SH_1, SH_i …, SH_n)

list (SH_1, SH_1, SH_i …, SH_n) : (#SH_1, #SH_i …, #SH_n)

CUI1	CUI2	Subheadings	#CUI1_CUI2

n=84

Co-occurence	CUI1	¬CUI1
CUI2	#CUI1_CUI2	#¬CUI1_CUI2
¬CUI2	#CUI1¬CUI2	#¬CUI1¬CUI2

Material and Methods

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

- Intermediate Occurrence Calculations (IMOC)
  - Step 2: Overall counts **OC**

    map (k1, v1) → list (OC, 1)

    reduce (OC, list (1, 1, …, 1)) → list (1, 1, …, 1) : #1

OC

  - Step 3: CUI1 counts #**CUI1**

    map (k1, v1) → list (CUI1, 1)

    reduce (CUI1, list (1, 1, …, 1)) → list (1, 1, …, 1) : #1

CUI1	#CUI1

  - Step 4: CUI2 counts #**CUI2**

    map (k1, v1) → list (CUI2, 1)

    reduce (CUI2, list (1, 1, …, 1)) → list (1, 1, …, 1) : #1

CUI2	#CUI2

Material and Methods

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

- Intermediate Occurrence Calculations (IMOC)
  - Step 5: Reduce Side Join on CUI1

CUI1	CUI2	Subheadings	#CUI1_CUI2

n=84

CUI1	#CUI1

#CUI1¬CUI2 = #CUI1 - #CUI1_CUI2

CUI1	CUI2	Subheadings	#CUI1_CUI2	#CUI1_¬CUI2

n=84

Co-occurence	CUI1	¬CUI1
CUI2	#CUI1_CUI2	#¬CUI1_CUI2
¬CUI2	#CUI1¬CUI2	#¬CUI1¬CUI2

Material and Methods

Step 1 ► Step 2 ► Step 3 ► Step 4 ► Step 5 ► Step 6

- Final Log-Likelihood Calculation (FLLC)
  - Step 6: Reduce Side Join on CUI2

CUI1	CUI2	Subheadings	#CUI1_CUI2	#CUI1_¬CUI2
		n=84		

CUI1	#CUI1

#¬CUI1_CUI2 = #CUI2 - #CUI1_CUI2

#¬CUI1¬CUI2 = OC - #CUI1_CUI2 - #CUI1¬CUI2 - #¬CUI1_CUI2

Material and Methods

Step 1 ► Step 2 ► Step 3 ► Step 4 ► Step 5 ► Step 6

Final Log-Likelihood Calculation (FLLC)

Co-occurence	CUI1	¬CUI1
CUI2	#CUI1_CUI2	#¬CUI1_CUI2
¬CUI2	#CUI1¬CUI2	#¬CUI1¬CUI2

$$H = -\sum_{i}^{n} p_i(log_b p_i)$$

$$LLR = 2\ (H(matrix) - H(rows) - H(cols))$$

H(matrix): Matrix entropy; H(rows): Sum of row entropies; H(cols): Sum of column entropies

CUI1	CUI2	Subheadings	#CUI1_CUI2	#CUI1_¬CUI2	#¬CUI1_CUI2	#¬CUI1_¬CUI2	LLR

n=84

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 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.

Slave instances	Calculation part			Sum
	IFAA	IMOC	FLLC	
2	50	27	36	113
4	29	13	17	59
10	16	9	7	32

- 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)

Unstructured Information Management Architecture
An Apache Project

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].

[1] Miñarro-Giménez, J. A., Kreuzthaler, M., & Schulz, S. (2015). Knowledge Extraction from MEDLINE by Combining Clustering with Natural Language Processing. In *AMIA Annual Symposium Proceedings* (Vol. 2015, p. 915). American Medical Informatics Association.

[2] Miñarro-Giménez, J. A., Kreuzthaler, M., Bernhardt-Melischnig, J., Martínez-Costa, C., & Schulz, S. (2014). Acquiring Plausible Predications from MEDLINE by Clustering MeSH Annotations. *Studies in health technology and informatics*, *216*, 716-720.

[3] Schulz, S., Costa, C. M., Kreuzthaler, M., Miñarro-Giménez, J. A., Andersen, U., Jensen, A. B., & Maegaard, B. (2014). Semantic relation discovery by using co-occurrence information. In *9th Language resources and evaluation conference (LREC)*.