Hasso Plattner Institut

IT Systems Engineering | Universität Potsdam

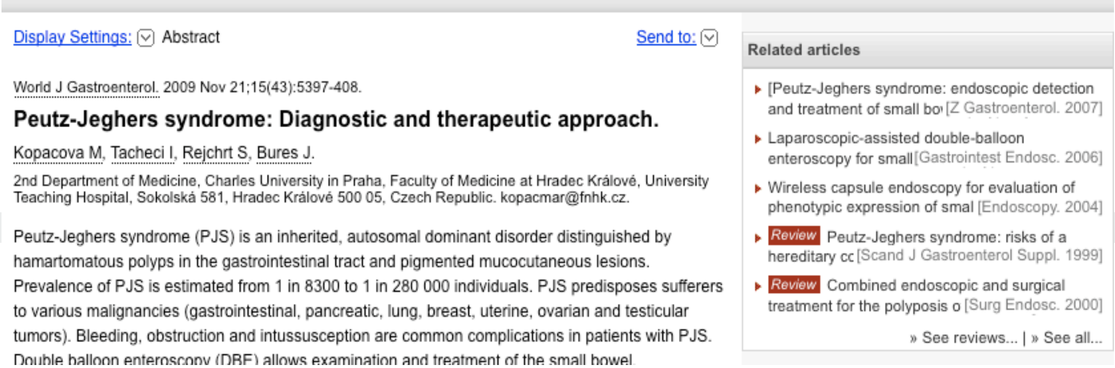
Brute Force and Indexed Approaches to Pairwise Document Similarity Comparisons with MapReduce

Jimmy Lin

Dandy Fenz
Thomas Berger

December, 15th 2009

First results



World J Gastroenterol. 2009 Nov 21;15(43):5397-408.

Peutz-Jeghers syndrome: Diagnostic and therapeutic approach.

Kopacova M, Tacheci I, Rejchrt S, Bures J.

2nd Department of Medicine, Charles University in Praha, Faculty of Medicine at Hradec Králové, University Teaching Hospital, Sokolská 581, Hradec Králové 500 05, Czech Republic. kopacmar@fnhk.cz.

Peutz-Jeghers syndrome (PJS) is an inherited, autosomal dominant disorder distinguished by hamartomatous polyps in the gastrointestinal tract and pigmented mucocutaneous lesions. Prevalence of PJS is estimated from 1 in 8300 to 1 in 280 000 individuals. PJS predisposes sufferers to various malignancies (gastrointestinal, pancreatic, lung, breast, uterine, ovarian and testicular tumors). Bleeding, obstruction and intussusception are common complications in patients with PJS. Double balloon enteroscopy (DBE) allows examination and treatment of the small bowel

Related articles

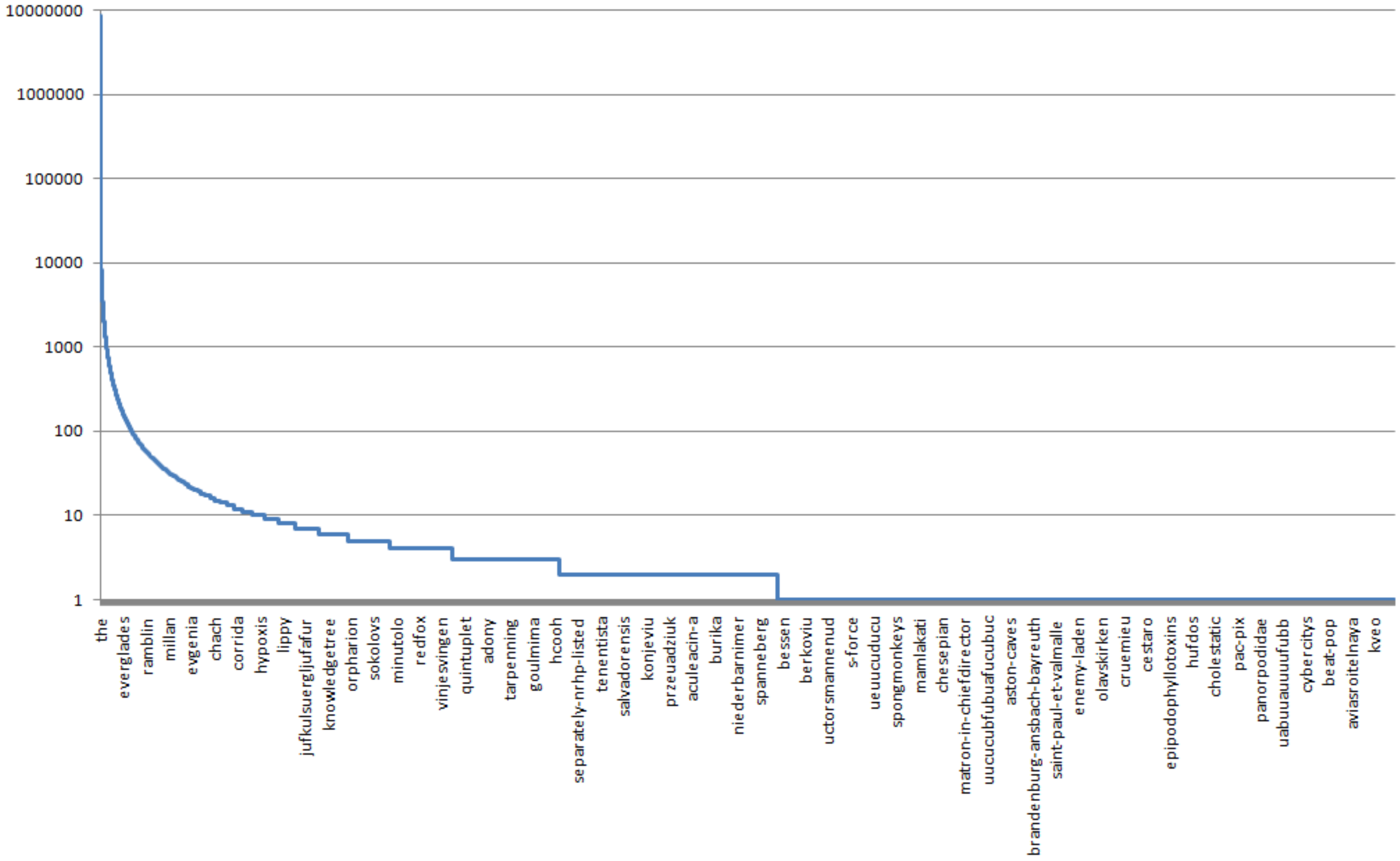
- [Peutz-Jeghers syndrome: endoscopic detection and treatment of small bo [Z Gastroenterol. 2007]
- Laparoscopic-assisted double-balloon enteroscopy for small [Gastrointest Endosc. 2006]
- Wireless capsule endoscopy for evaluation of phenotypic expression of smal [Endoscopy. 2004]
- Review Peutz-Jeghers syndrome: risks of a hereditary cc [Scand J Gastroenterol Suppl. 1999]
- Review Combined endoscopic and surgical treatment for the polyposis o [Surg Endosc. 2000]

» See reviews... | » See all...

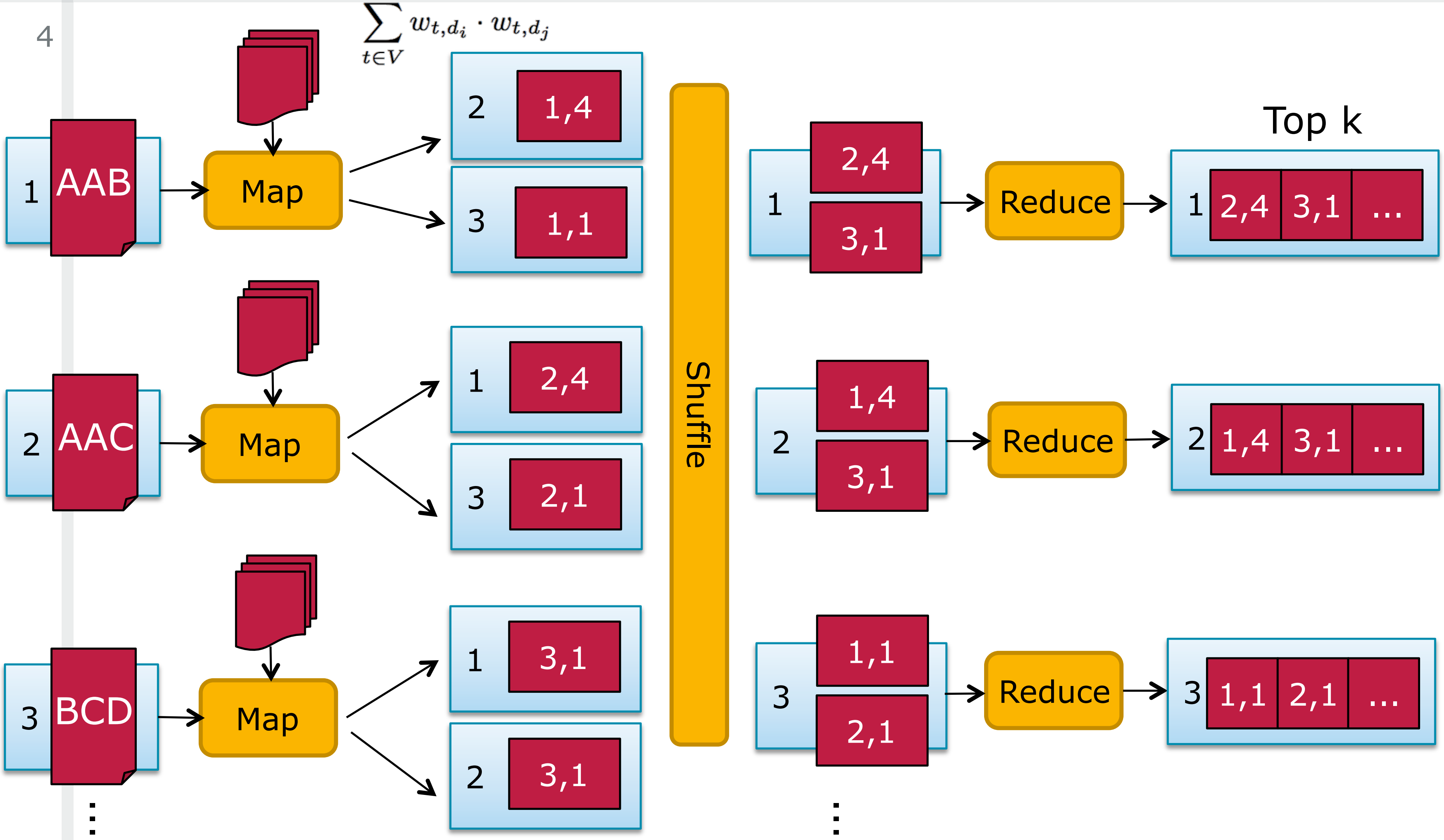
- Given situation:
  - □ Database with wikipedia abstracts
  - □ Task: Show „more-like-this" articles

- Implementation of following tasks:
  - □ Preparation of input data ✓
  - □ Brute Force algorithm ✓
  - □ Inverted index creation ✓
  - □ Postings Cartesian Product (PCP) algorithm ✓
  - □ Prove PCP efficiency
  - □ PCP approxima

Our collection: frequency of terms

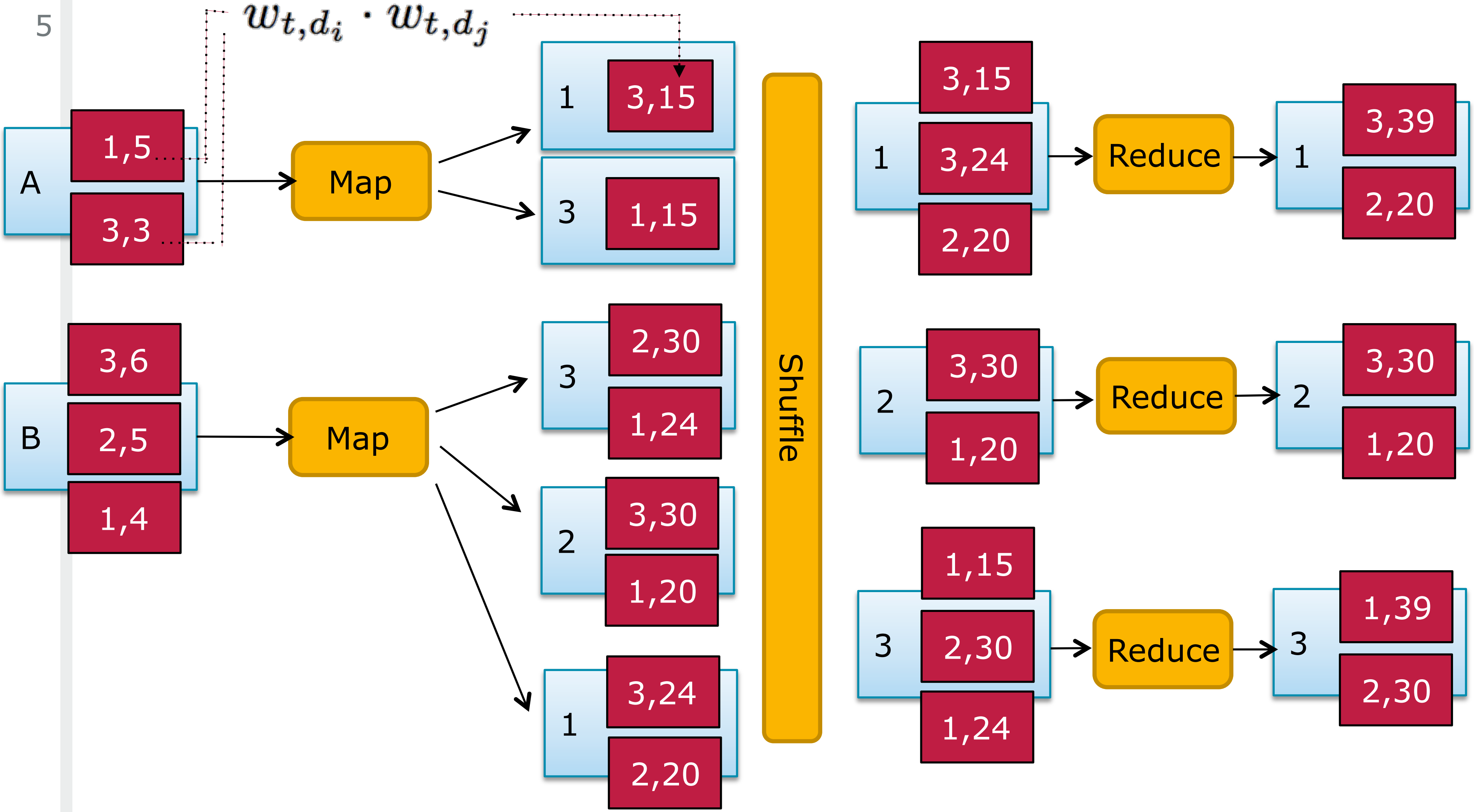


Brute Force (BF)



$$\sum_{t \in V} w_{t,d_i} \cdot w_{t,d_j}$$

Postings Cartesian Product (PCP)



Comments on Hadoop

- Framework relatively easy to configure and understand

- Documentation slightly outdated
  - Plenty of example code now marked as deprecated
  - Hints for current version rare

- Complex data types implementations
  - Unfortunately not provided by Hadoop Framework
  - Found useful implementation on author‘s website:
    - http://www.umiacs.umd.edu/~jimmylin/cloud9/docs/index.html

Next steps

- General optimizations
  - Efficient usage of data structures

- Use approximations
  - Implementation of proposed approx. for PCP
  - Research for alternative aproaches

- Generation of statistics
  - Compare efficiency of different algorithms
  - Analysis of results
    - Is comparison of wikipedia abstracts reasonable?