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

November, 17th 2009

Introduction

Display Settings: Abstract Send to:

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

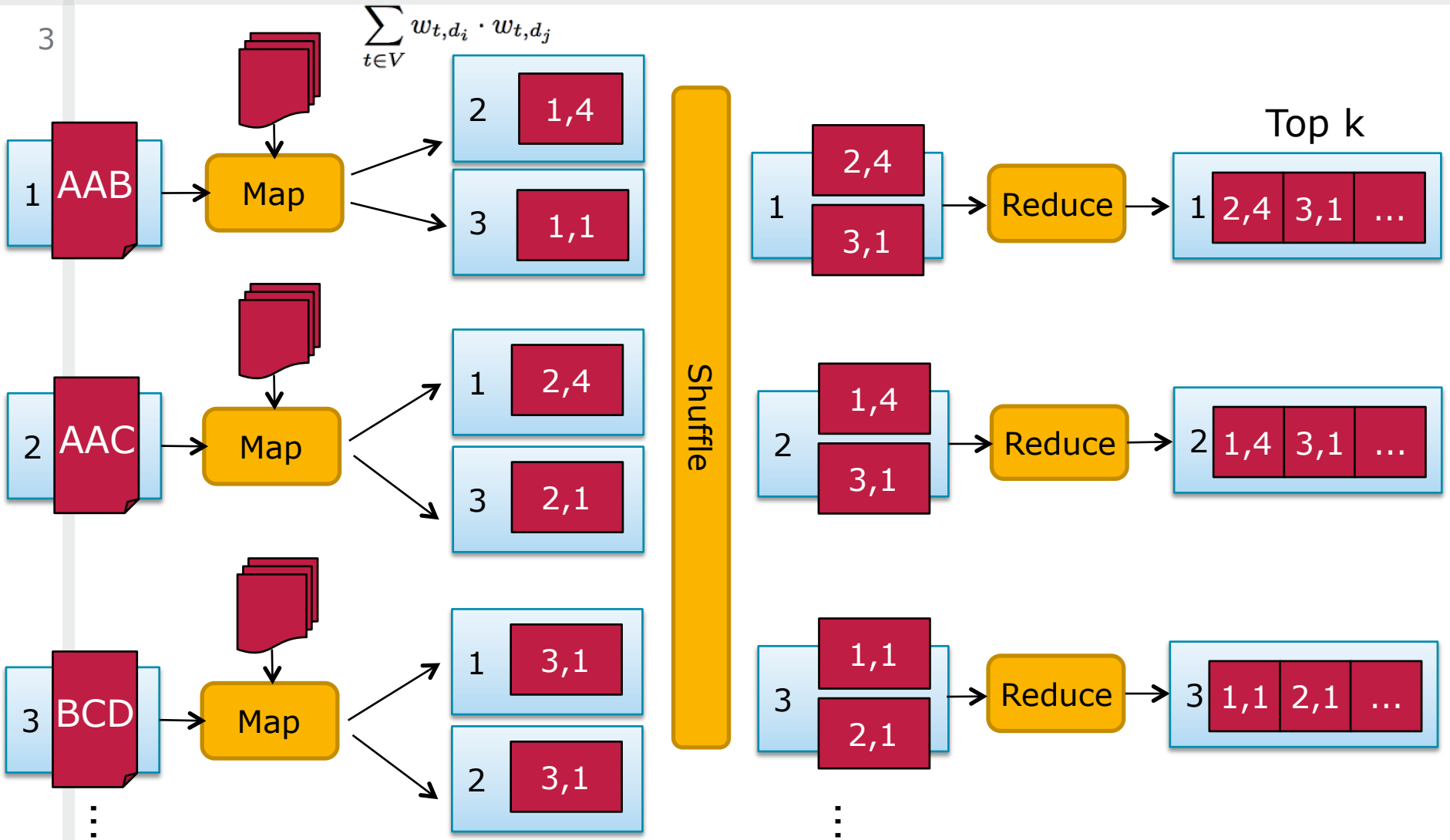
Related articles

- [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]

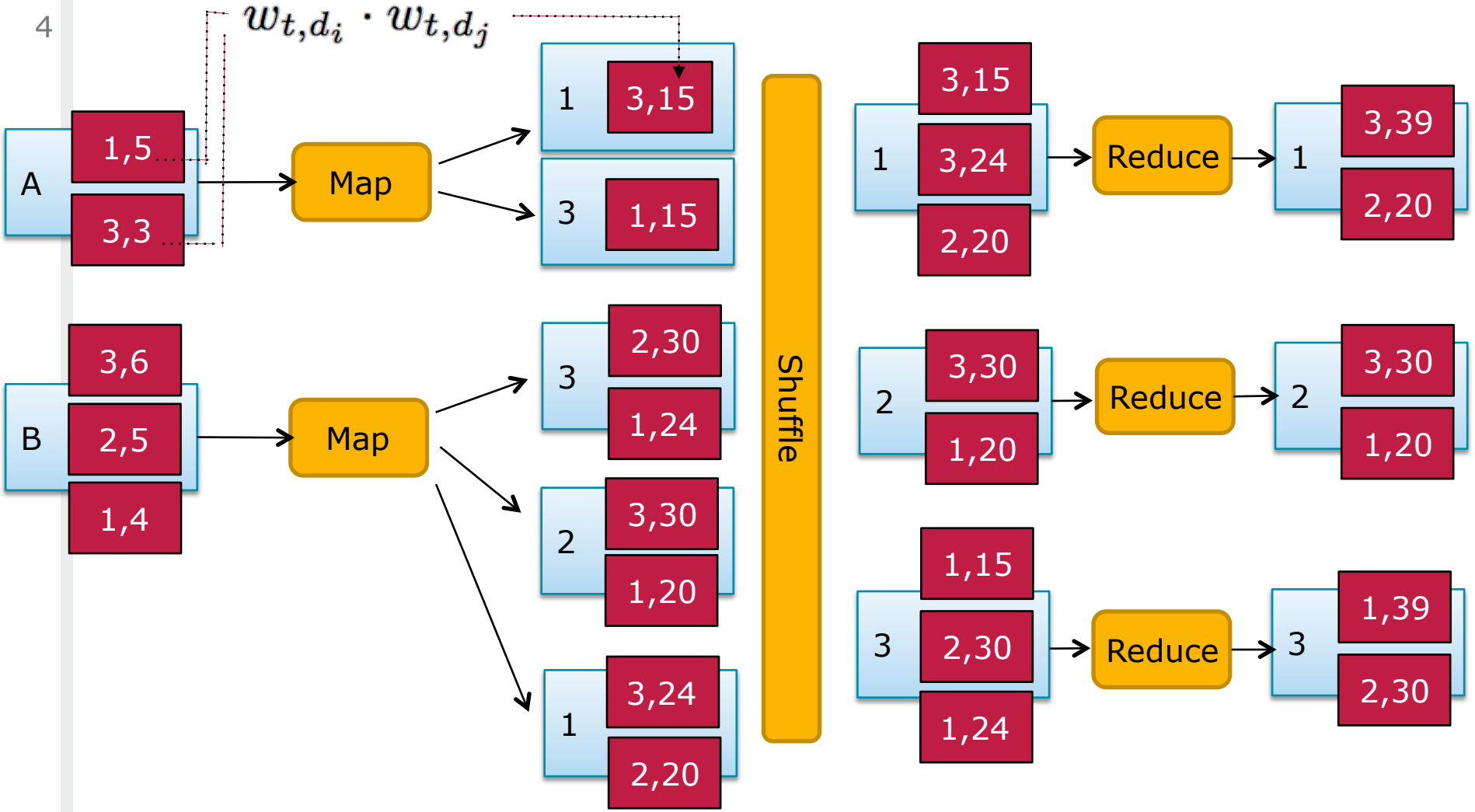
» See reviews... | » See all...

- Given situation:
  - Database with life-sciences articles (PUBMED)
  - Task: Show „more-like-this“ articles
- Map&Reduce based solutions:
  - Brute Force (BF)
  - Using an inverted index
    - Parallel Query (PQ)
    - Postings Cartesian Product (PCP)
- Optimizations:
  - Three approaches based on approximations

Brute Force (BF)

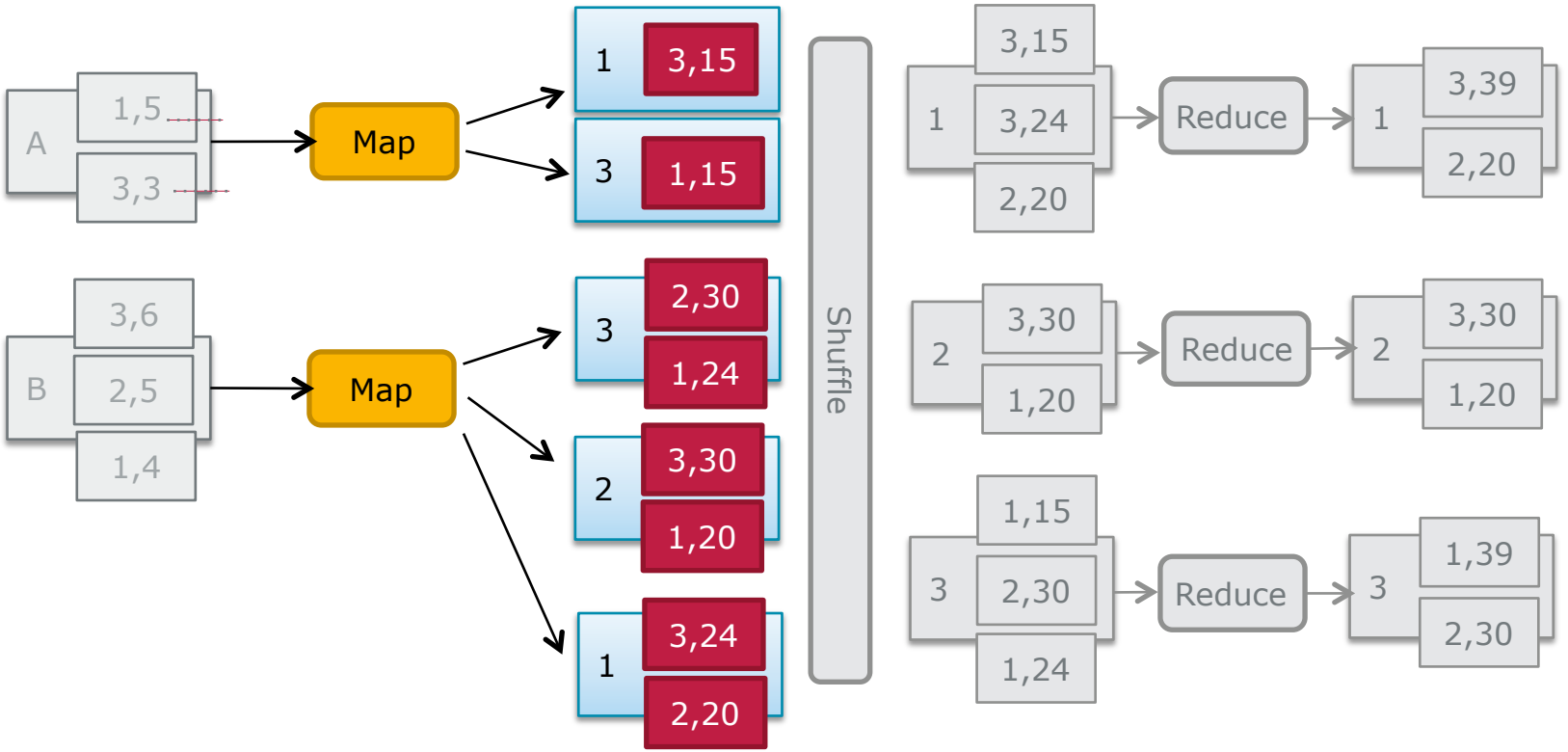


Postings Cartesian Product (PCP)



Approximation A)
Accumulator size limit

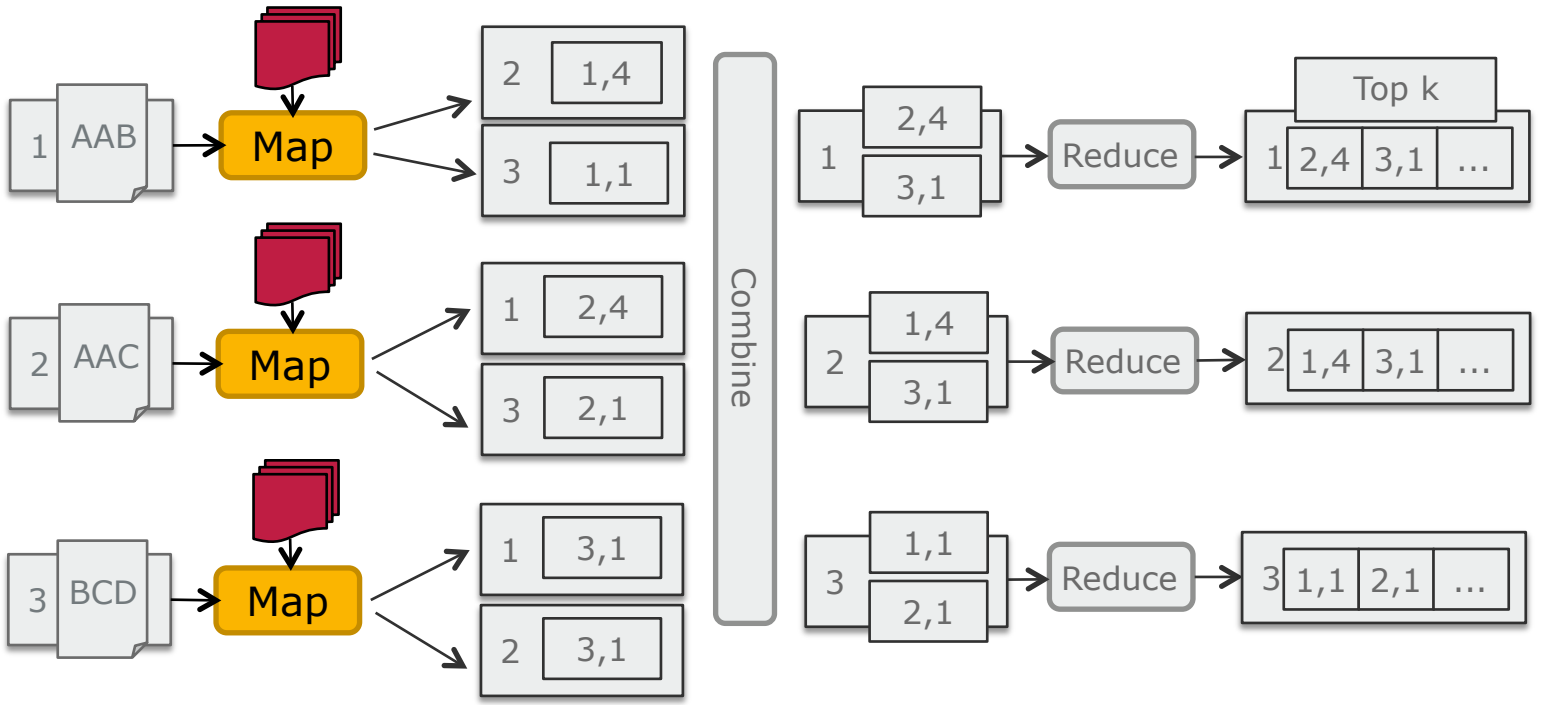
- Simply limit size of associative array
- Applicable for index-based approaches (PCP and PQ)



PCP

Approximation B)
Term limit

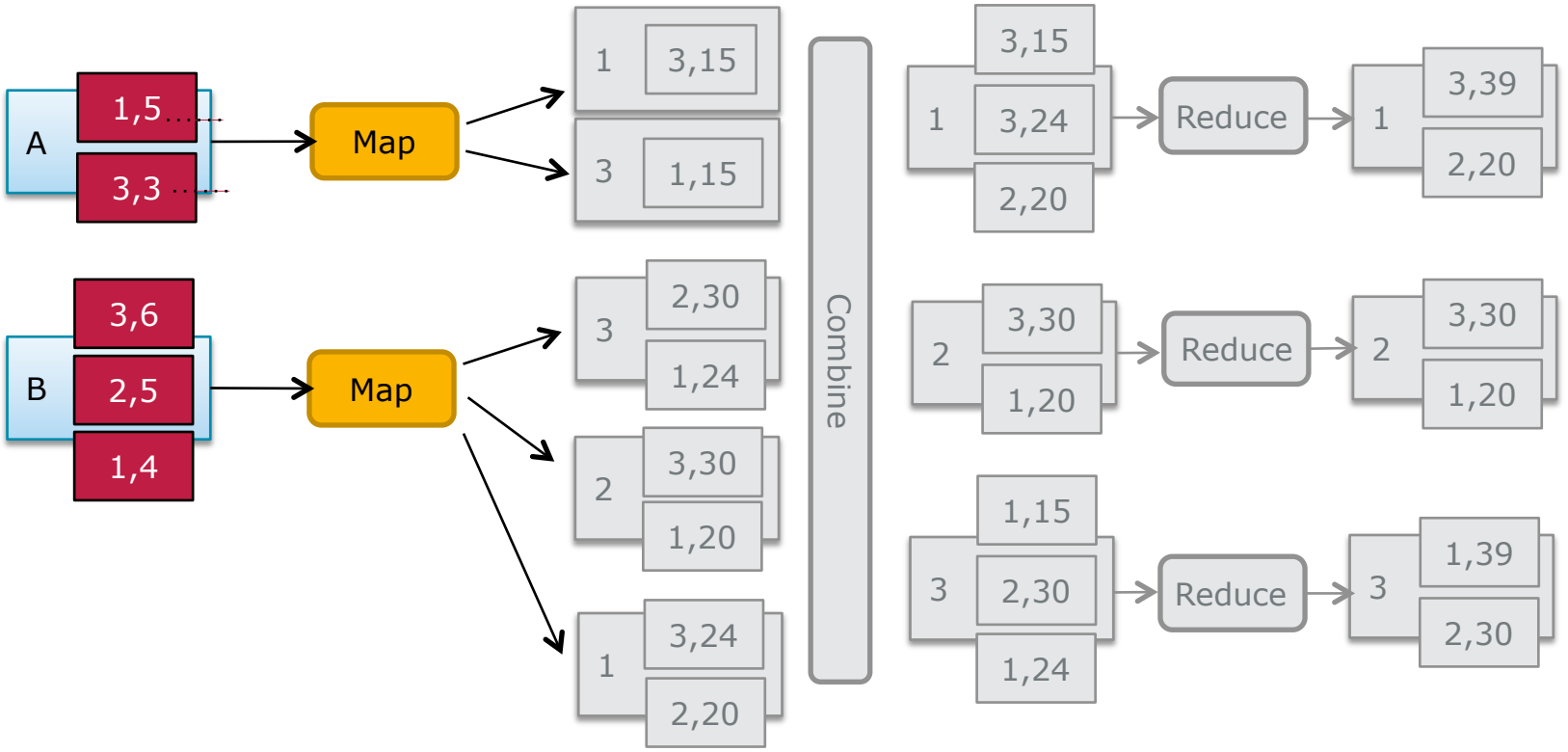
- Only the n terms with lowest document frequencies (df) are loaded into the mapper.
- Precondition: Access to all documents -> BF and PQ approach



BF

Approximation C)
Document-frequency limit

- Ignore posting when df for that term is higher than threshold
- Applicable for index-based approaches (PCP and PQ)



PCP

Outlook & Remarks

- Implementation of following algorithms:
  - Brute Force
  - Inverted index creation & Postings Cartesian Product
  - Prove PCP efficiency
- Our dataset: Wikipedia abstracts (DBPedia)

- Remarks on the paper:
  - Differences between algorithms not well-formulated
  - Concepts such as „blocks“ not scientificly defined
  - Experimental setting not clearly described