

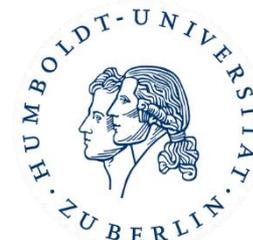
Textual Similarity Search on Big Data

OCT 12TH, 2016

Prof. Johann-Christoph Freytag, Ph.D.,
Fabian Fier

{headtcentre}

humboldt-elsevier
advanced data & text centre



About Us



Research Integrity

Similarity Search



Christoph



Fabian

Motivation 1: Plagiarism Detection

Copying in suspect text (not reference text) with no reference	Grayscale color
Significant copying: multiple identical sentences in a paragraph	6
some copying: ≥ 5 contiguous words within a sentence of ≥ 5 words	5
near copying: multiple (≥ 3) exact phrases (≥ 3 words) overlap in contiguous sentences in a paragraph	4
similarity 1: several (< 3) exact phrases (≥ 3 words) overlap in contiguous sentences in a paragraph or more than 9 words in sequence	3
similarity 2: many (≥ 5) of the exact same words (excluding function words) in contiguous sentences in the same paragraph	2
similarity 2: topic overlap with facts and standard phrases	1
Standard words or phrases	0
General knowledge / facts	0

Motivation 2: Enhancing Web Search

trekking travel

Web Shopping Bilder News Maps Mehr ▾ Suchoptionen

Ungefähr 38.400.000 Ergebnisse (0,44 Sekunden)

Trekking Reisen - aktivferien.com
Anzeige www.aktivferien.com/ ▾
Aktiver Urlaub für aktive Menschen an den Traumplätzen dieser Welt.
Anden in Peru - Kilimanjaro - Himalaya - Ecuador mit Galapagos

Trek Travel Luxury Cycling Vacations of a Lifetime
trektravel.com/ ▾ Diese Seite übersetzen
Creating cycling experiences of a lifetime around the world. Join us for a bike tour you will remember. Includes the use of an award winning **Trek Bicycle**.

Trekking Travel
www.trekkingtravel.com.vn/ ▾ Diese Seite übersetzen
Sapa **trekking**, homestay **travel** tours, Halong bay, **Trekking** Yen Tu mountain, Truc Lam Zen monastery, Hanoi, Hue, Hoi-an, Hochiminh, Mekong delta...
Open Bus - Payment - Saigon Tours & Around - Hotels in Hai Phong

Trekking Travel Expediciones, Cabalgatas / Horseback Riding
www.trekking-travel.com.ar/ ▾ Diese Seite übersetzen
Offers horseback riding and **trekking** in the Los Andes Mountain Range and mountaineering expeditions to Aconcagua. Includes itineraries and schedule.
Cruce de Los Andes a caballo - Nuestra empresa - 2 días, Cabalgata "Villavicencio"

Walking & Trekking Holidays - Explore
<https://www.explore.co.uk/walking-and-trekking-to...> ▾ Diese Seite übersetzen
Walking and **Trekking** with Explore We specialise in small group **trekking** and walking holidays. On these you will be in MrZen **Travel** Website Development.



Nepal!

Motivation 2: Enhancing Web Search

Home Destinations Themes About Deals				
Everest Base Camp				♥ Add
15 Oct 2016	29 Oct 2016		Fully booked	+
17 Oct 2016	31 Oct 2016	OR	Available	€1,445 +
19 Oct 2016	2 Nov 2016	OR	1 space left!	€1,445 +
22 Oct 2016	5 Nov 2016	OR	Available	€1,445 +
26 Oct 2016	9 Nov 2016	OR	4 spaces left!	€1,445 +
29 Oct 2016	12 Nov 2016	OR	Available	€1,445 +
2 Nov 2016	16 Nov 2016		Fully booked	+
9 Nov 2016	23 Nov 2016	DG	Available	€1,445 +
12 Nov 2016	26 Nov 2016	DG	Available	€1,445 +
16 Nov 2016	30 Nov 2016	DG	Available	€1,445 +
19 Nov 2016	3 Dec 2016	DG	Available	€1,445 +

Web search is **only textual**

no **similarity search**

SELECT * FROM web WHERE
text SIMILAR TO CURRENT website

Problem Statement

Input:

- Set R of documents
- Search document s
- Similarity measure $sim(r, s) \mapsto [0; 1]$
- Similarity threshold θ

Output:

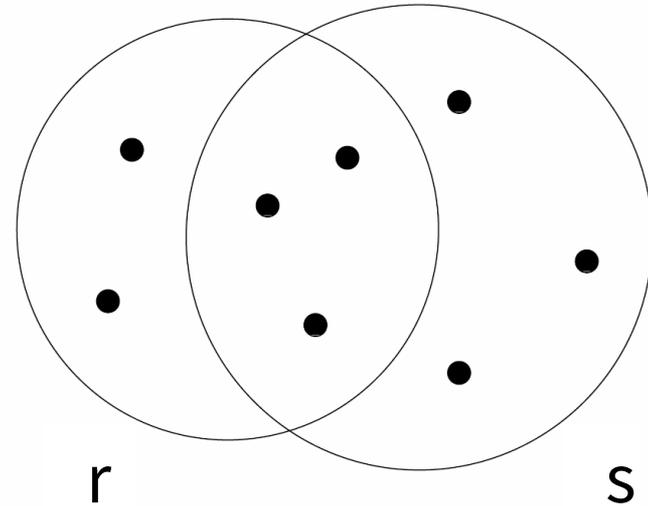
- $\{r \in R \mid sim(r, s) \geq \theta\}$

Example: Jaccard Similarity

$r = \{\text{lorem ipsum consetetur sadipscing elitr}\}$

$s = \{\text{magna aliquyam erat consetetur sadipscing elitr}\}$

$$\text{sim}(r, s) = \frac{|r \cap s|}{|r \cup s|} = \frac{3}{8}$$



Naive Approach

Given: search document s , input dataset R , similarity threshold θ

For each $r \in R$:

 If $\text{sim}(r, s) \geq \theta$:

 Output r

Feasibility?

Our Approach

- Use distributed similarity **join** algorithms (MapReduce)
 - Similar problem: find **all pairs** of similar documents
 - $\{(r, s) \in (R \times R) \mid \text{sim}(r, s) \geq \theta\}$
 - Use intermediate results for indexing
- Find the best join approach(es) and use them for search
 - Metric partitioning
 - Filter-and-verify

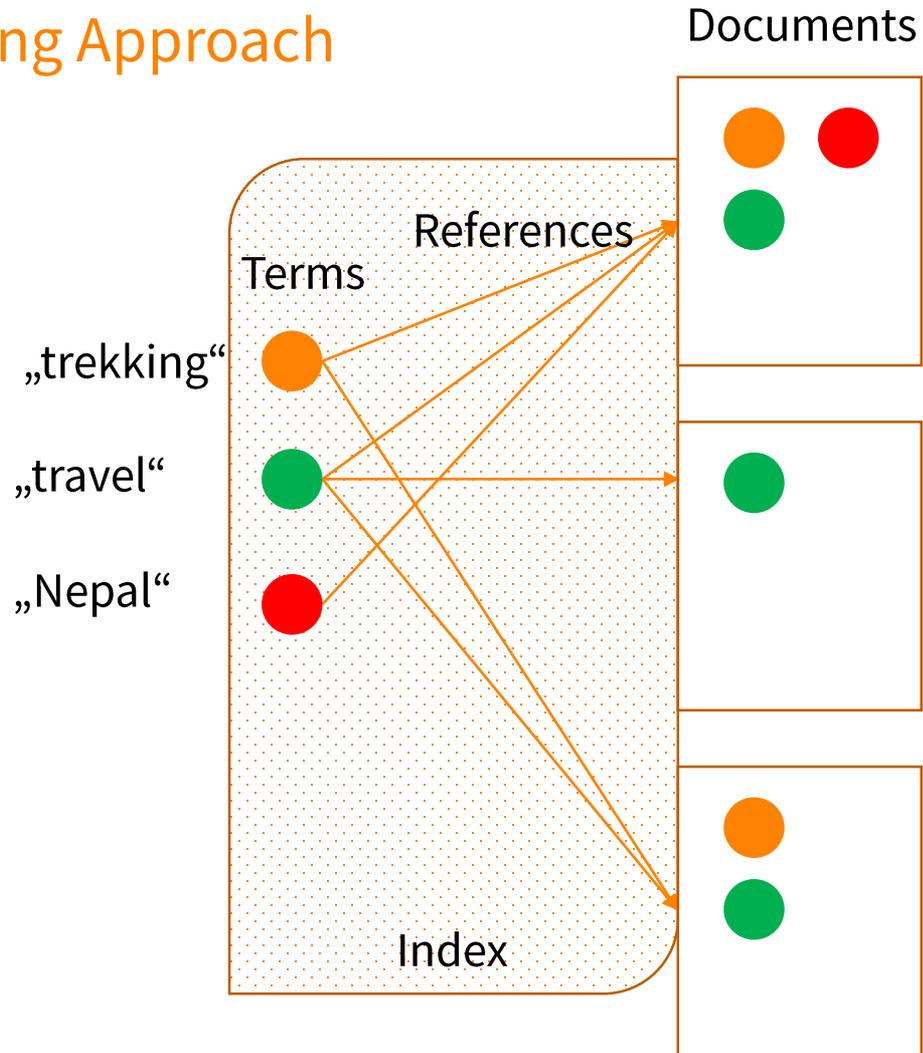
Filter-and-Verify Approach

- Generate a set of **candidate pairs**:
 - superset of result set
 - use of filters: orders of magnitude smaller than cross product (usually)
- For each candidate pair, verify if it meets the similarity threshold

Filter-and-Verify: Full-Filtering Approach

- Build inverted index $\{\langle term, \{document\}\rangle\}$
- For each element $\langle term, \{document\}\rangle$ of the inverted index:
 - For each document pair (r, s) :
 - If $\text{sim}(r, s) \geq \theta$:
 - Output $\langle r, s \rangle$

Specific problems?



Filter-and-Verify: Prefix Filtering

Example:

$r = \{\text{Lorem ipsum dolor sit } \underline{\text{consetetur}} \text{ sadipscing elit}\}$

$s = \{\underline{\text{magna aliquyam erat }} \underline{\text{consetetur}} \text{ sadipscing elit}\}$

$$\text{prefix}_{\text{Jaccard}}(r, 0.5) = 5$$

$$\text{prefix}_{\text{Jaccard}}(s, 0.5) = 4$$

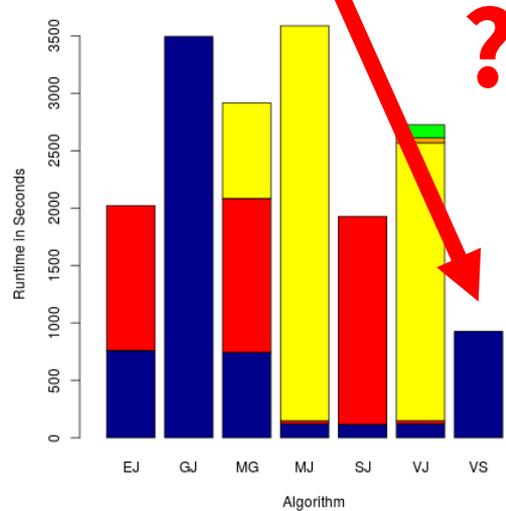
Build inverted index only over the prefix: result stays complete.

Optimization: use global token order. Sort each set/document in ascending order (already true for this example).

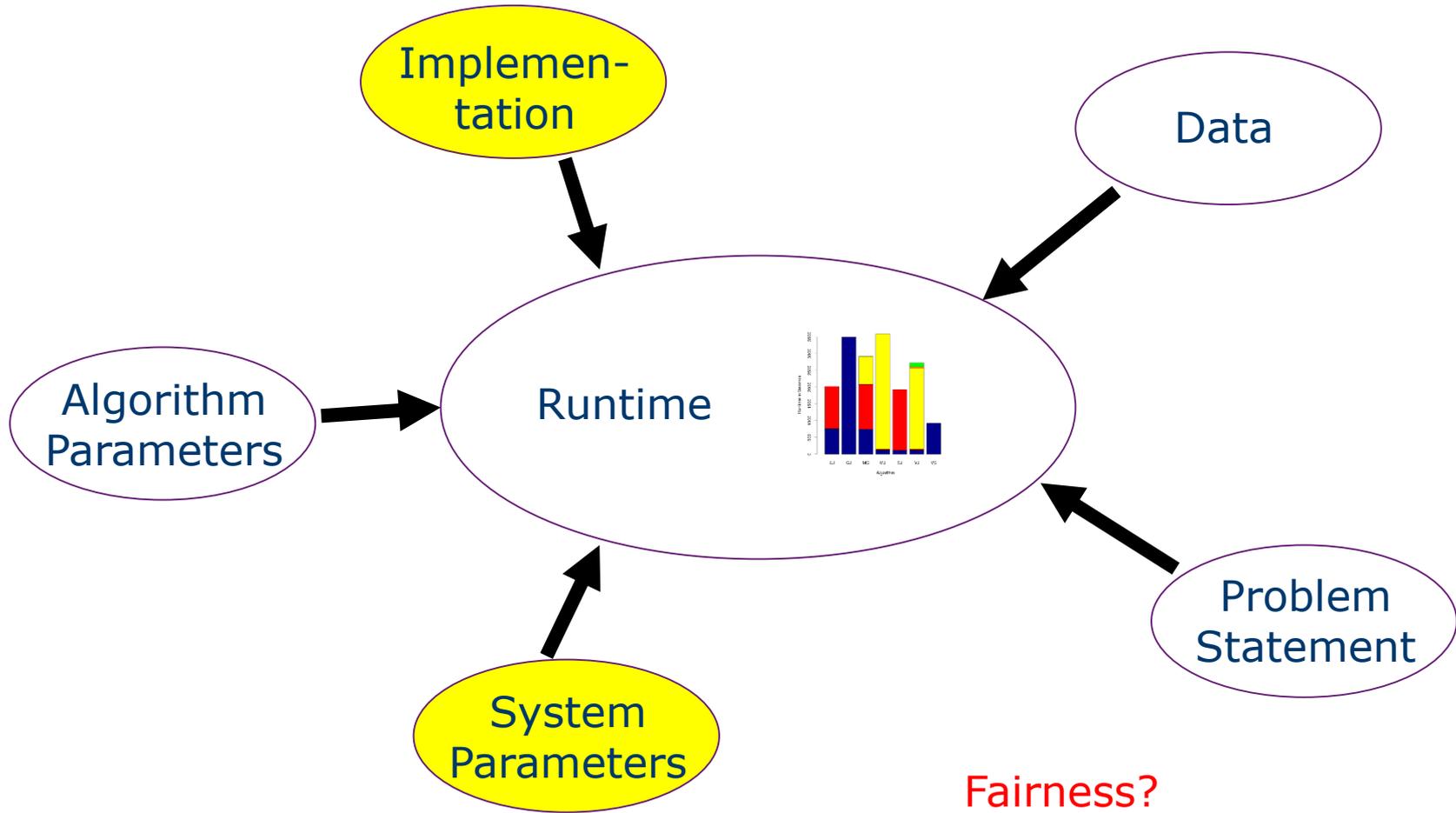
Evaluation

- We experimentally compare the runtime of MapReduce join algorithms

The *V-SMART-Join* algorithms are very efficient and scalable in the number of entities, as well as their cardinalities. They were **up to 30 times faster** than the state of the art algorithm, *VCL*, when compared on a real dataset of a small size.



Comparability Issue



Fairness?

Comparability Issue

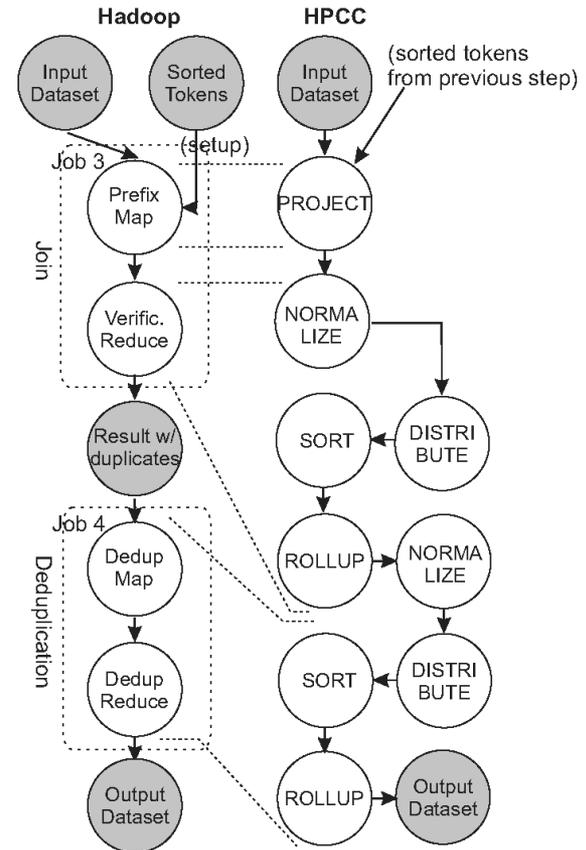
- Implementation
 - Combine
 - Prune locally
- System Parameters
 - Data Split
 - Parallelization
 - Memory

Idea: Use a MapReduce implementation with less tuning knobs than Hadoop: HPCC subsumes MapReduce

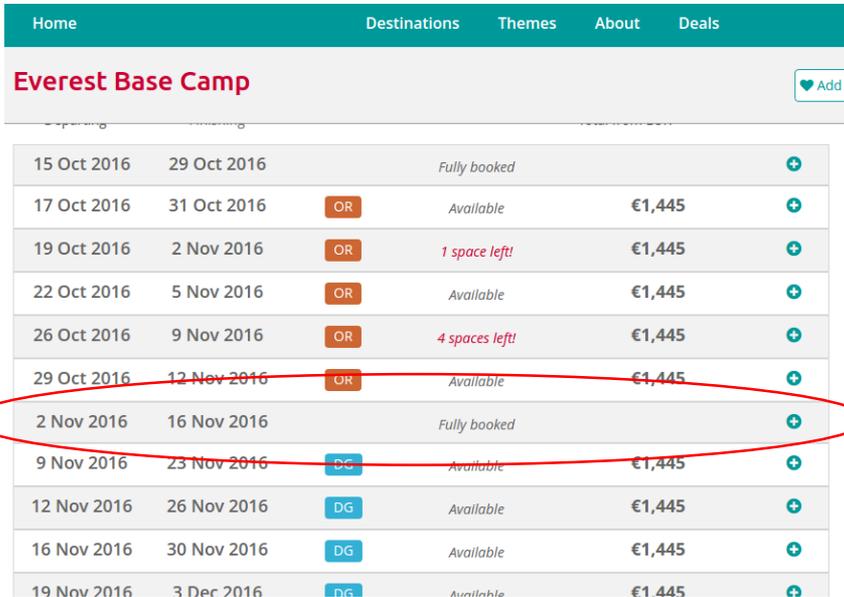


Comparability Issue

- Implementation
 - Combine - same
 - Prune locally - same
- System Parameters
 - Data Split – good default
 - Parallelization – good default
 - Memory - ?



Future: Enhancing Web Search



Home Destinations Themes About Deals				
Everest Base Camp ♥ Add				
15 Oct 2016	29 Oct 2016		Fully booked	+
17 Oct 2016	31 Oct 2016	OR	Available	€1,445 +
19 Oct 2016	2 Nov 2016	OR	1 space left!	€1,445 +
22 Oct 2016	5 Nov 2016	OR	Available	€1,445 +
26 Oct 2016	9 Nov 2016	OR	4 spaces left!	€1,445 +
29 Oct 2016	12 Nov 2016	OR	Available	€1,445 +
2 Nov 2016	16 Nov 2016		Fully booked	+
9 Nov 2016	23 Nov 2016	DG	Available	€1,445 +
12 Nov 2016	26 Nov 2016	DG	Available	€1,445 +
16 Nov 2016	30 Nov 2016	DG	Available	€1,445 +
19 Nov 2016	3 Dec 2016	DG	Available	€1,445 +

Web search is **only textual**

no **similarity search**

```
SELECT * FROM web WHERE  
text SIMILAR TO CURRENT website  
AND start_date > 11/01/2016  
AND end_date < 11/30/2016  
AND availability > 1  
AND location CLOSE TO Nepal
```

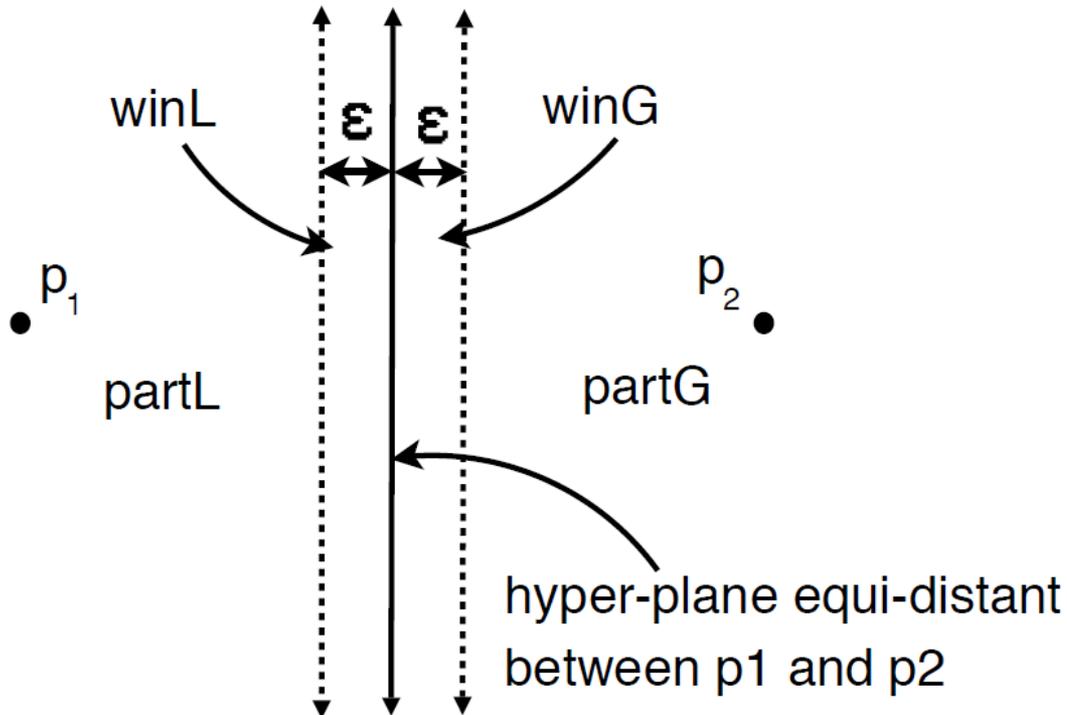
Thank you!

Questions?

Backup

Metric Partitioning Approach

- Idea: use similarity measure to iteratively partition the space until partitions fit into main memory



p_1, p_2 : random pivot elements
 $\epsilon = 1 - \theta$

(From Jacox and Samet:
Metric Space Similarity
Joins)