ZigZag: Supporting Similarity Queries on Vector Space Models

Wenhai Li, Lingfeng Deng, Yang Li
(Wuhan University, China)

Chen Li (UC Irvine)

SIGMOD 2018, Houston, TX

Example: HR manager looking for candidates

Skills:

» Art

» Stats

» DB

» Alice
  > Art (Average)
  > Writing (Good)
  > Stats (Excellent)
  > Programming (Good)

» Bob
  > Writing (Excellent)
  > Stats (Excellent)
  > Programming (Good)
  > DB (Good)

Similarity

Global token weight ("IDF")

Skill	Global Weight
Art	1
Writing	2
Stats	2
Progamming	4
DB	3

Record-specific token degree ("TF")

Skill	Employee, Degree	Employee, Degree
Art	Alice, 1	
Writing	Alice, 2	Bob, 3
Stats	Alice, 3	Bob, 3
Progamming	Alice, 2	Bob, 2
DB		Bob, 2

Search condition

Token degree in record R

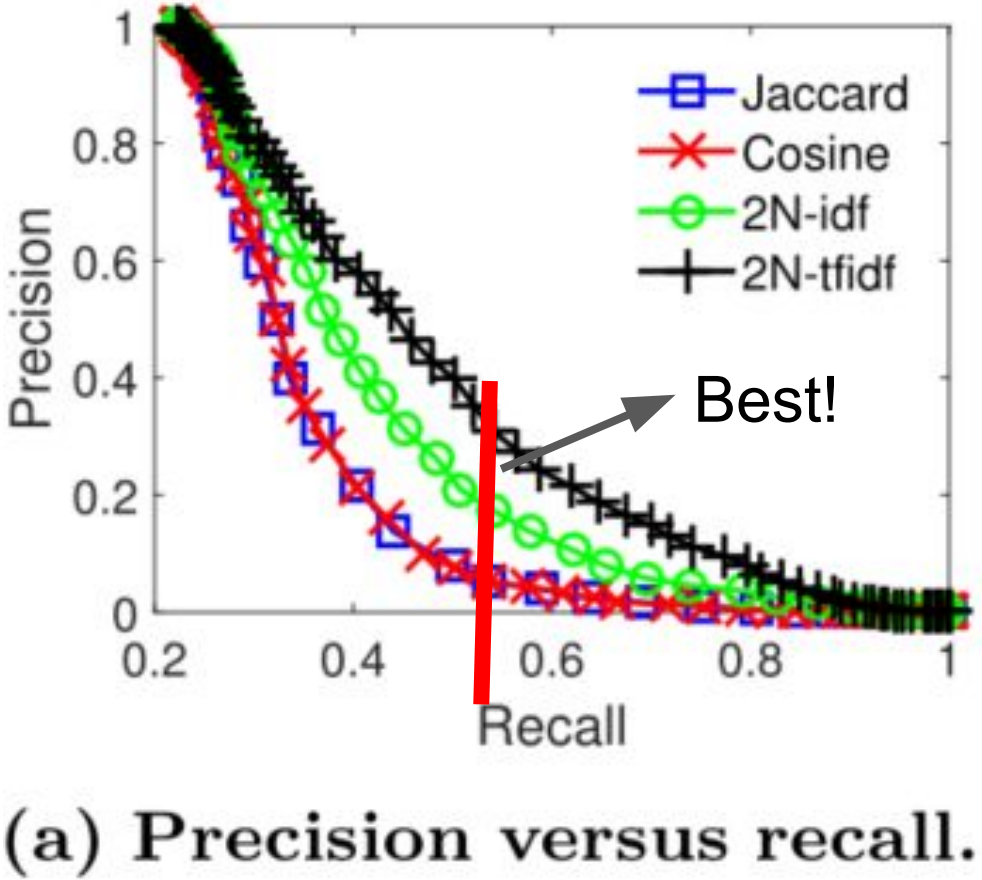
Token weight

$$\frac{\sum_{t \in R \cap Q} f(t,R)w(t)f(t,Q)w(t)}{\sqrt{\sum_{t \in R} (f(t,R)w(t))^2}\sqrt{\sum_{t \in Q} (f(t,Q)w(t))^2}} \geq \tau$$

Length of record R

Wenhai Li, et al, (Wuhan Uni.), **Chen Li (UC Irvine)**

Similarity Benefits

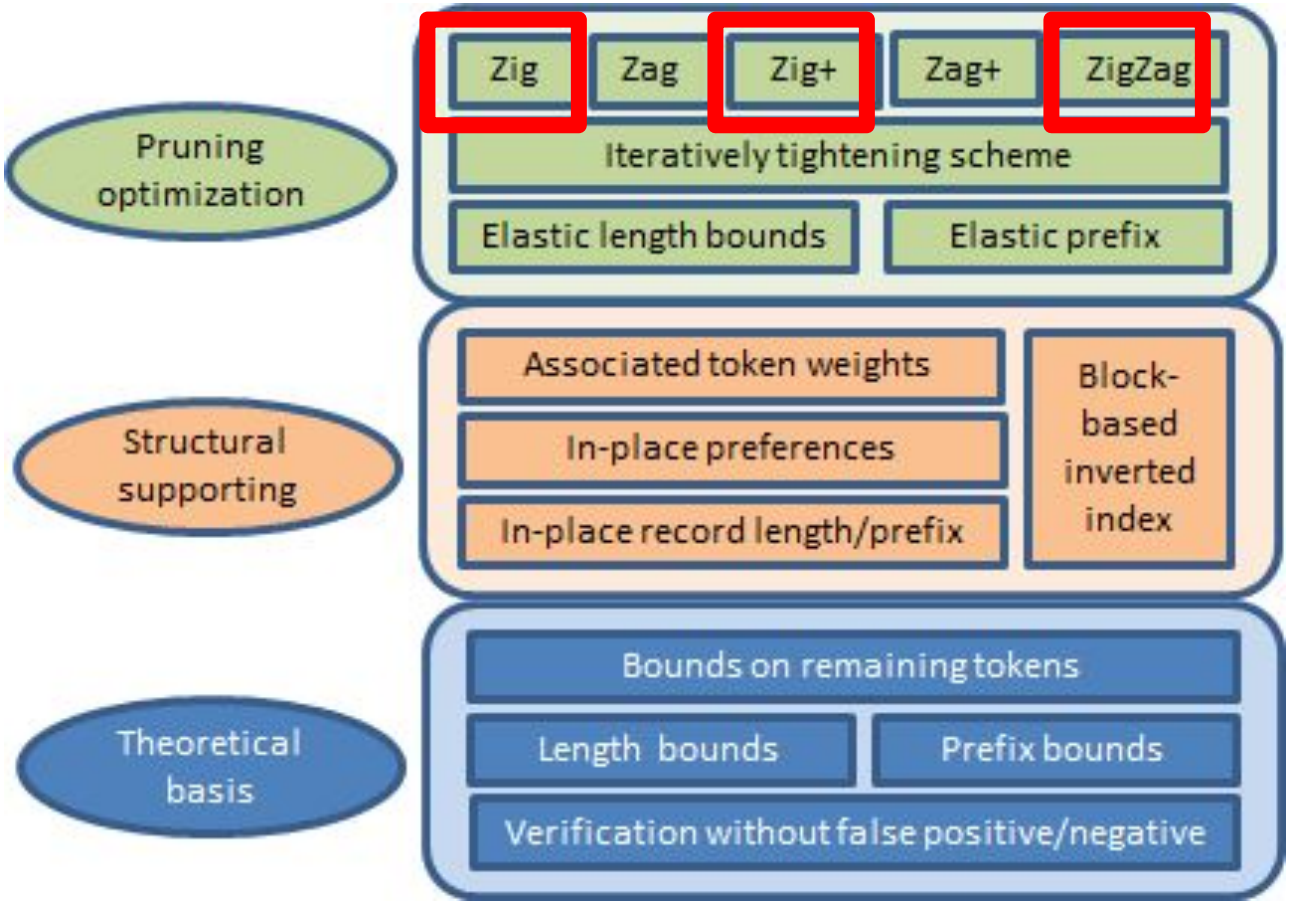


(a) Precision versus recall.

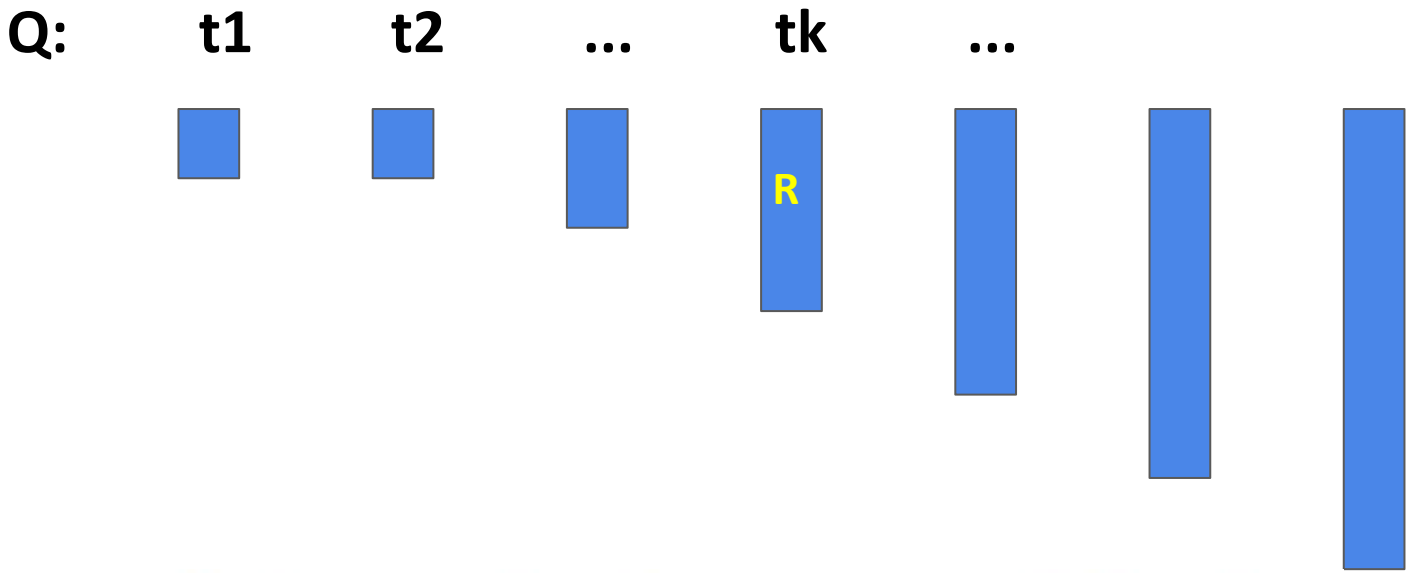
Wenhai Li, et al, (Wuhan Uni.), **Chen Li (UC Irvine)**

Contributions

- » Indexing
- » Algorithms using bounds
- » Tightening bounds iteratively

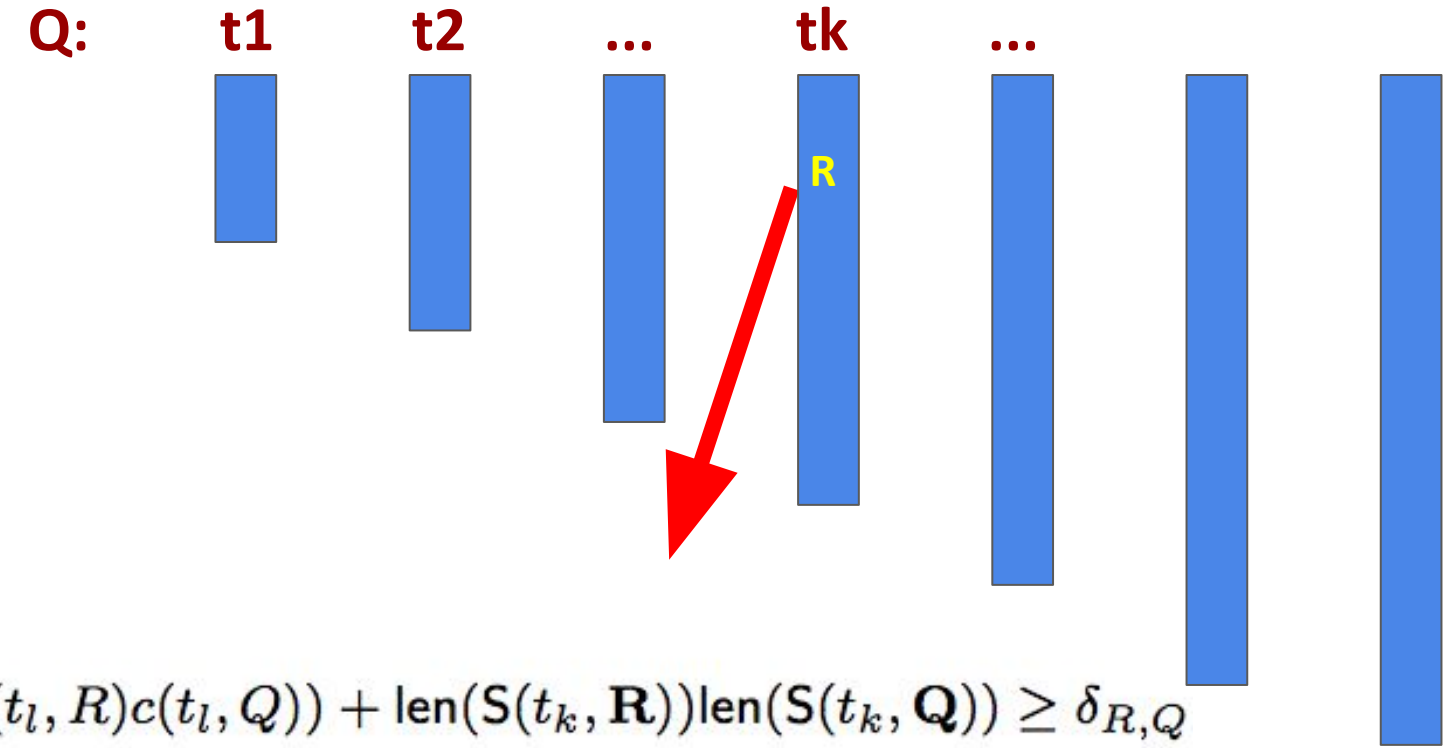


Naive algorithm: scan lists one by one



$$\sum_{t \in R \cap Q} c(t, R)c(t, Q) \geq \tau \mathsf{len}(R)\mathsf{len}(Q).$$

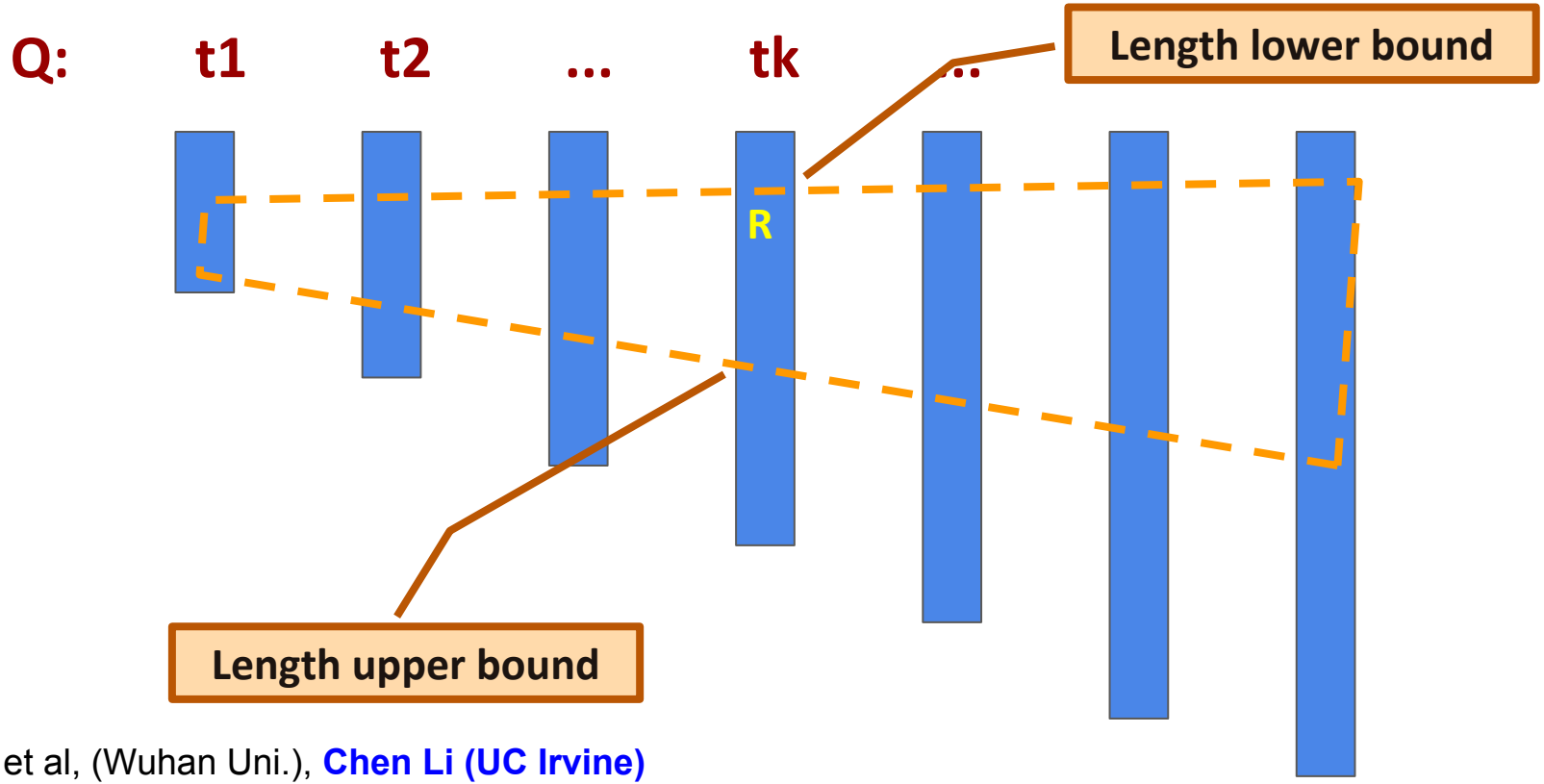
Pruning using length bound of remaining tokens



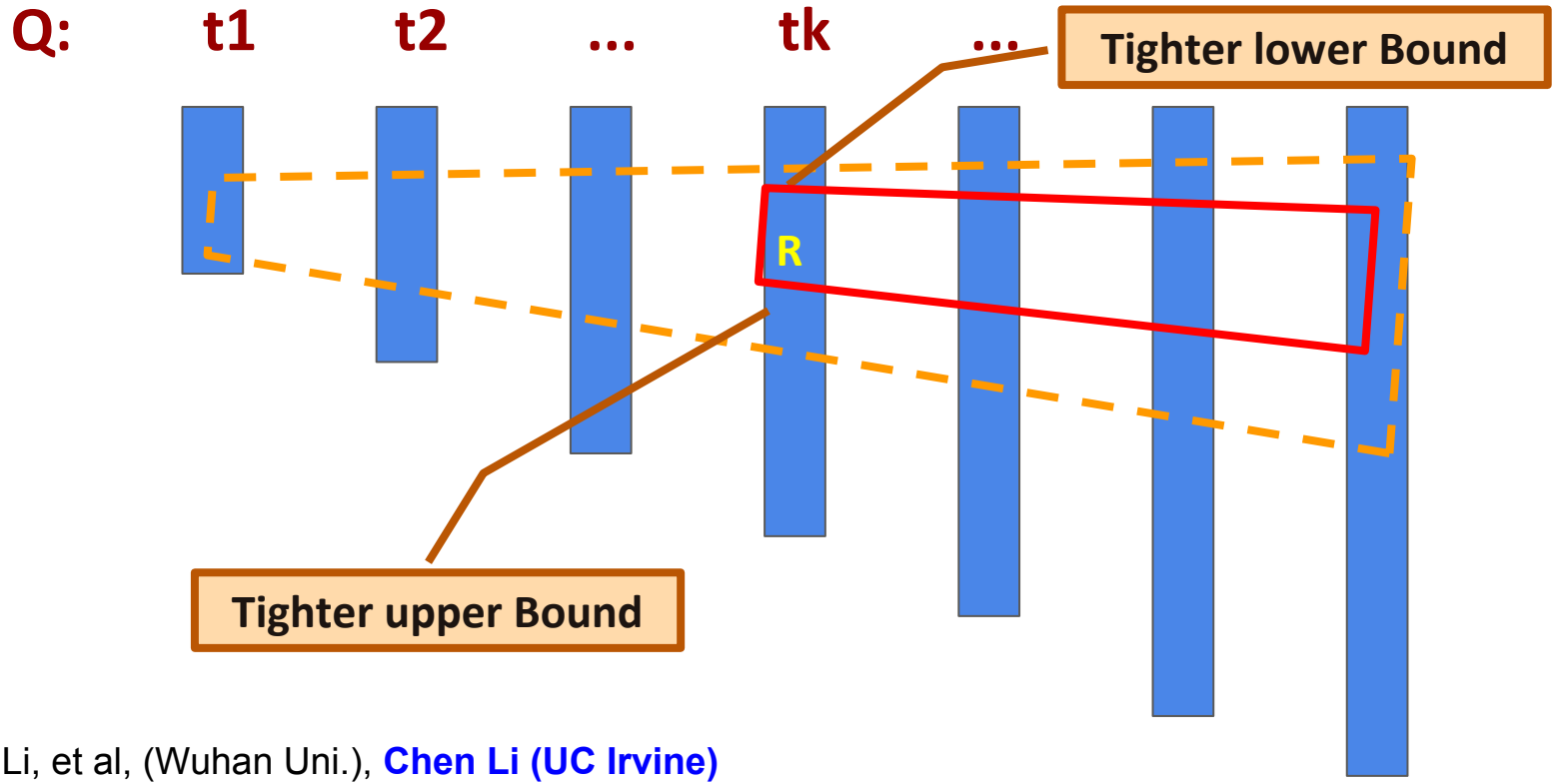
$$\sum_{l=1}^{k} (c(t_l, R)c(t_l, Q)) + \text{len}(\text{S}(t_k, \mathbf{R}))\text{len}(\text{S}(t_k, \mathbf{Q})) \geq \delta_{R,Q}$$

Wenhai Li, et al, (Wuhan Uni.), **Chen Li (UC Irvine)**

Pruning using length bounds per list (“Zig”)



Tightening length bounds per list (“Zig+”)

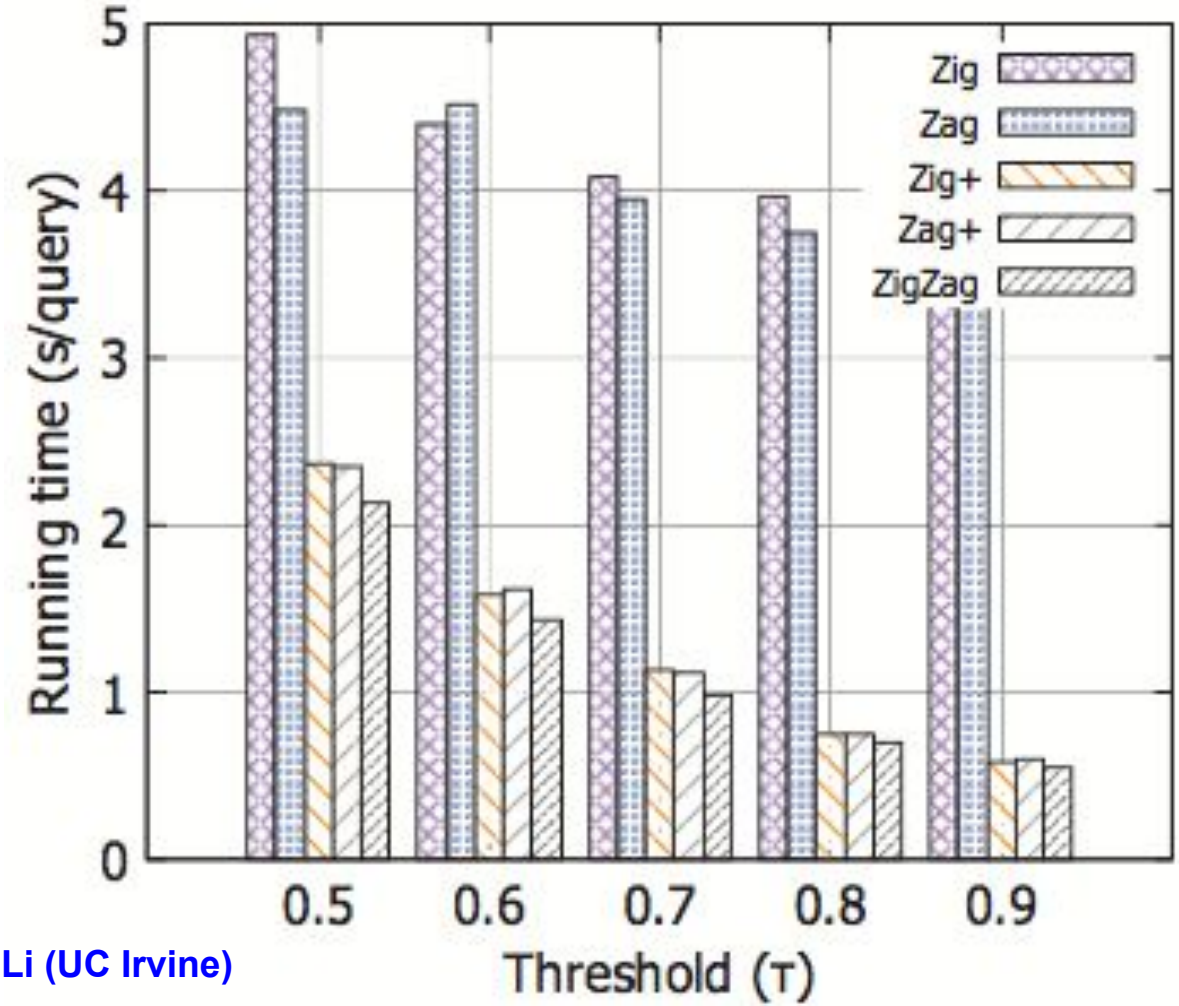


Other results

- A bound on the prefix (“Zag” and “Zag+”)
- Tightening prefix iteratively (“ZigZag”)

Experiments

- 3 data sets
- +30M records each
- In-memory
- Hard disks
- SSD



Wenhai Li, et al, (Wuhan Uni.), **Chen Li (UC Irvine)**

ZigZag: Supporting Similarity Queries on Vector Space Models

Wenhai Li, Lingfeng Deng, Yang Li
(Wuhan University, China)

Chen Li (UC Irvine)

Thank you!