Trie-based Edit Similarity Search & Join [SSS&J Workshop]

Jianbin Qin, Xiaoling Zhou, **Wei Wang**

University of New South Wales, Australia

Chuan Xiao

Nagoya University, Japan

Outline

- Context
- Problem Definition & Motivations
- Overview of Our Approach
- Conclusions

NB: Many other approaches not covered here !

Context

sub-parts	Fixed Length	Variable Length
Overlapping	q-gram ed-join [VLDB08]	vgram [Li et al, VLDB 07] NGPP [SIGMOD11]
Non-overlapping	q-gram-chunk [SIGMOD11, TODS] PASS-JOIN [Li et al, VLDB 12]	Vchunk [TKDE12]

- Similarity query processing proj
  - Exact similarity query processing
    - Edit distance-related
      - Non-trie-based methods
      - Trie-based method
        - Error-tolerant prefix matching [VLDB13, LEVA]
        - Extension to error-tolerant exact match (aka. edit similarity search) in [SSS&J 2013]

Only targets at Geonames [search & join]

Problem Definition

- With an edit distance threshold t in $[0, t_{max}]$
- ed-search(Q, $\mathfrak{S}$) = { S in $\mathfrak{S}$ | ed(S, Q) ≤ t }

- ed-join($\mathfrak{R}$, $\mathfrak{S}$) = {<R, S> | ed(R, S) ≤ t, R in $\mathfrak{R}$, S in $\mathfrak{S}$ }
  - Special case: self ed-join

- Comments
  - The workshop specification is slightly different
    - Allow t = 0
    - Ed Join: output <x, y> and <y, x>; output <x, x>; input not sorted
    - Ed Search: queries with different t; queries given in batch; pretty generous constraints in indexing time& size.

Source: Hadjieleftheriou & Li, VLDB09 tutorial

Motivations /1

UCSD

Yannis Papakonstantinou

Case Western

Meral Ozsoyoglu

AT&T--Research

Marios Hadjieleftheriou

http://www.informatik.uni-trier.de/~ley/db/indices/a-tree/index.html

Source: http://www.ics.uci.edu/~chenli/pubs.html

Motivations /2

- Typographical errors
  - Why everybdoy can undrstand this?
  - Person's names (or other Named entities)

4. **Efficient Approximate Search on String Collections (Tutorial),** Marios Hadjeleftheriou and Chen Li, VLDB 2009. [PDF], [Part I], [Part II].
5. **Efficient Approximate Search on String Collections (Tutorial),** Marios Hadjieleftheriou, Chen Li, ICDE 2009, [PPT-Part1], [PPT-part2].
6. **Quality-Aware Retrieval of Data Objects from Autonomous Sources for Web-Based Repositories,** Houtan Shirani-Mehr, Chen Li, Gang Liang, Michal Shmueli-Scheuer, ICDE 2008 (poster). [PDF]
7. **Communication-Efficient Query Answering with Quality Guarantees in Client-Server**

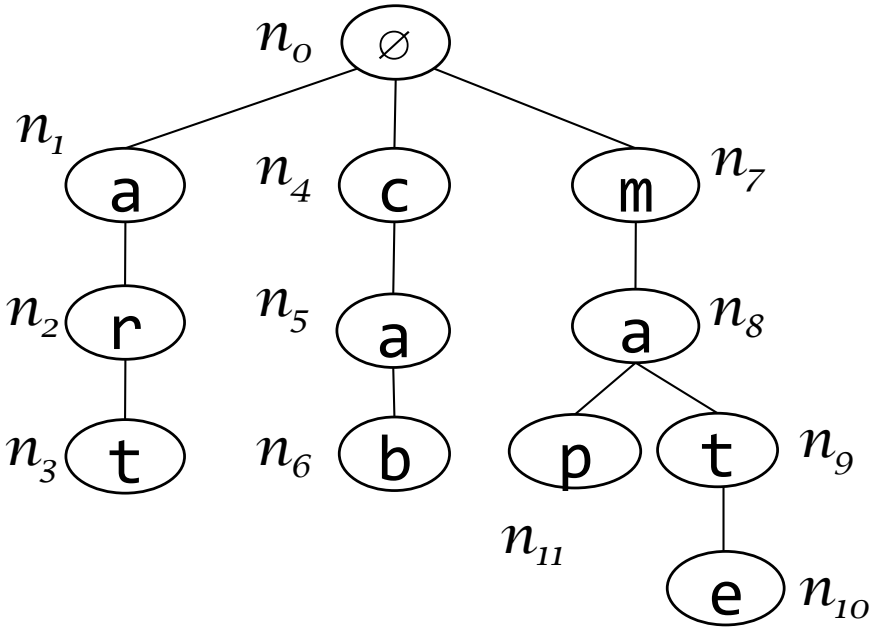
Motivations /3

- Big data intel project Department of Defense, Australia
  - Cross-document Coreference Resolution (CDCR)
    - Requires finding highly similar "mentions" based on a sophisticate similarity measure, which includes edit distance (to measure orthographic similarity)
    - Naïve solution requires $O(n^2)$ comparisons, where n = 40.3 million in a recent study [Singh et al, ACL HLT 2011]
    - (Self) Similarity join can help

Trie-based Ed-Search [Chaudhuri & Kaushik, SIGMOD09, Ji et al, WWW09, Li et al, VLDBJ11]

Generalization of t=0

- Idea:
  - Incrementally maintain the Active Node Sets (ANS) for each query prefix Q[1..i]
  - ANS = {trie node n | ed(n, Q[1..i]) ≤ t }

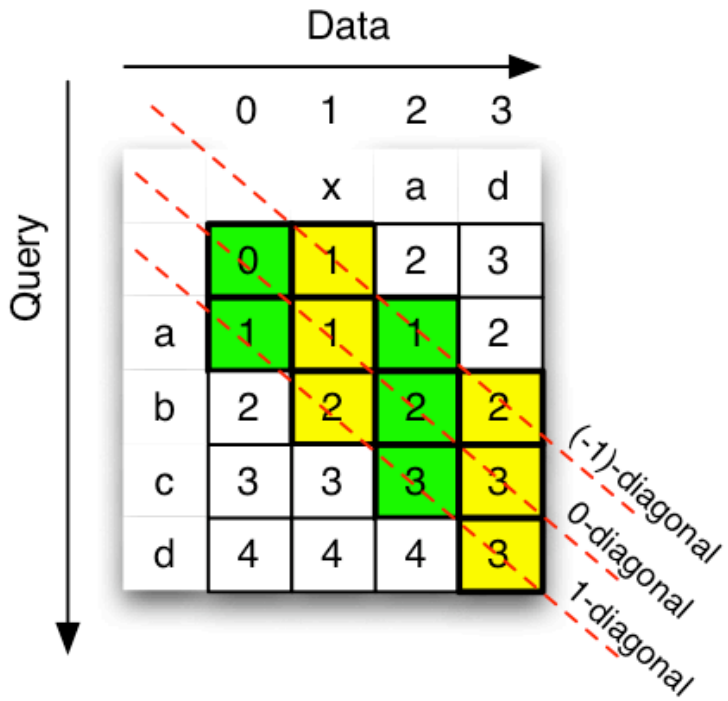


Step	Query	Active States & Their Edit Distances
1	∅	$\{n_0,0\},\{n_1,1\},\{n_4,1\},\{n_7,1\}$
2	c	$\{n_0,1\},\{n_1,1\},\{n_4,0\},\{n_5,1\},\{n_7,1\}$
3	ca	$\{n_1,1\},\{n_4,1\},\{n_5,0\},\{n_6,1\},\{n_8,1\}$
4	cat	$\{n_5,1\},\{n_6,1\},\{n_9,1\}$

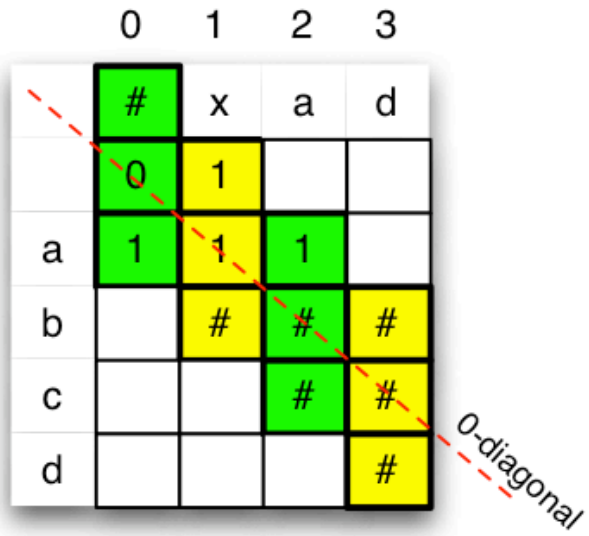
Q = cat, t = 1

Improvements

- EVA
- LEVA
- Adaptations



(a) Edit Distance Matrix and Diagonals



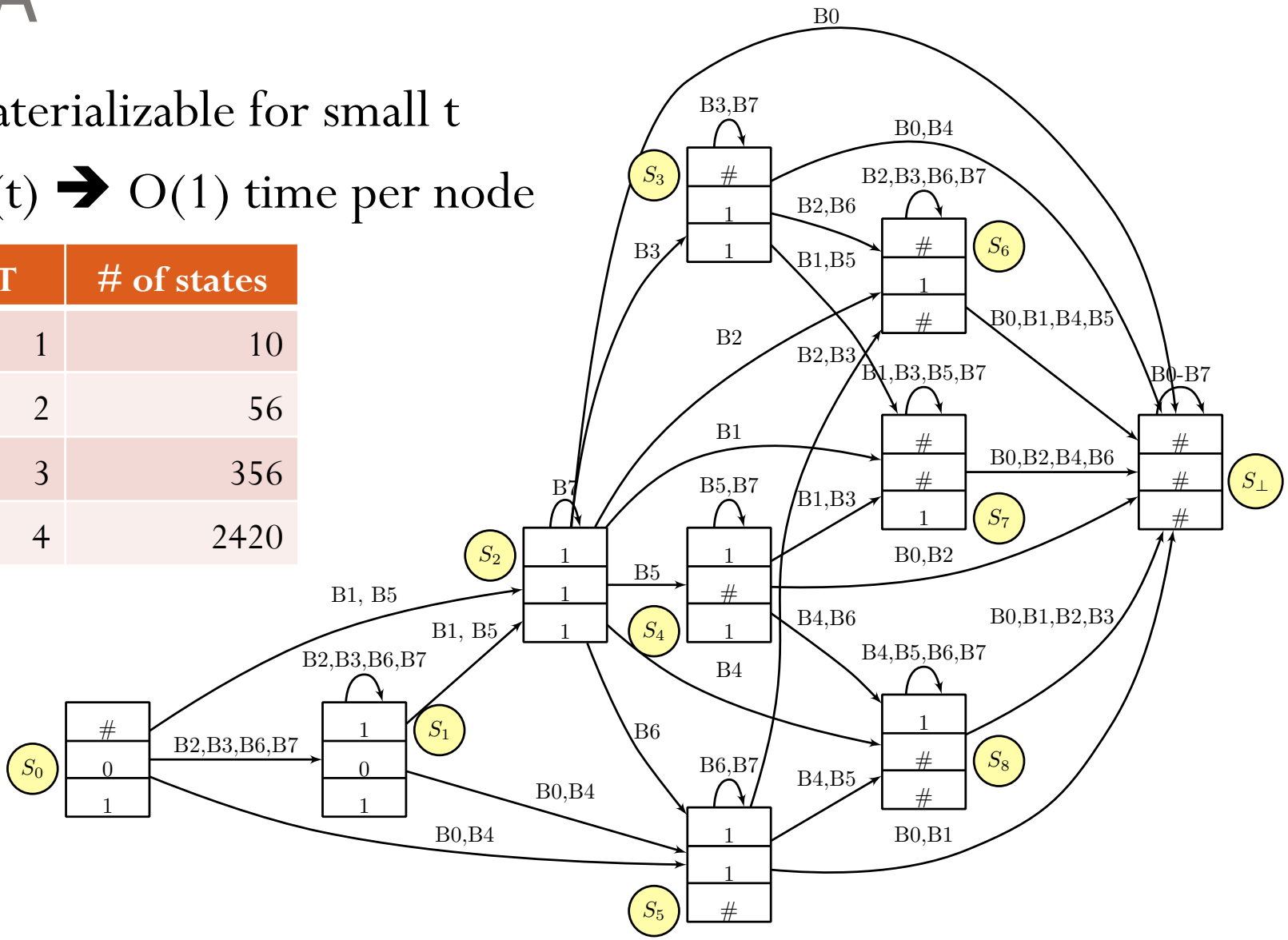
(b) Edit Vectors (τ = 1)

EVA

- Materializable for small t
- O(t) ➔ O(1) time per node

T	# of states
1	10
2	56
3	356
4	2420

EVA for t = 1

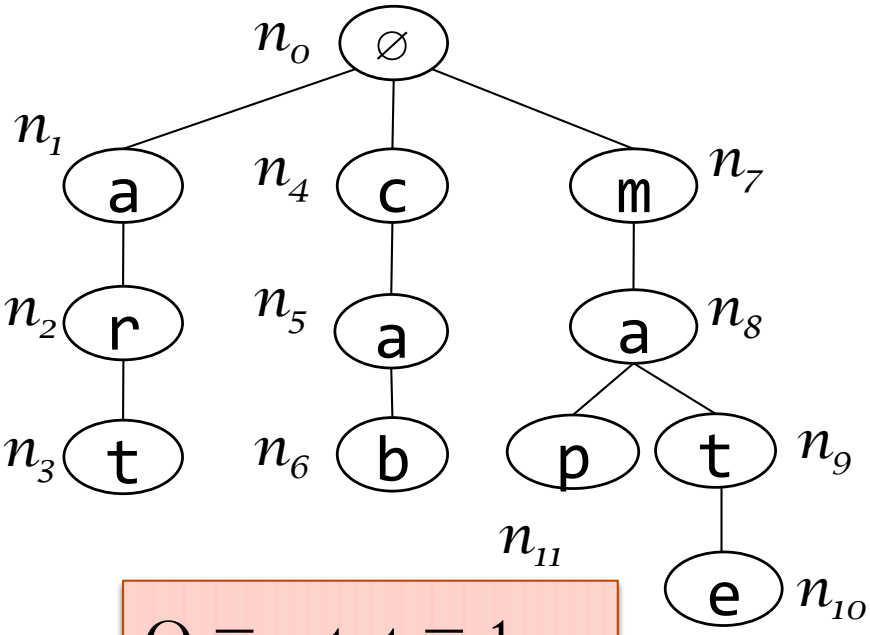


LEVA

- Maintain only potentially feasible nodes

Step	Query	Active States & Their Edit Distances
1	∅	$\{n_0,0\},\{n_1,1\},\{n_4,1\},\{n_7,1\}$
2	c	$\{n_0,1\},\{n_1,1\},\{n_4,0\},\{n_5,1\},\{n_7,1\}$
3	ca	$\{n_1,1\},\{n_4,1\},\{n_5,0\},\{n_6,1\},\{n_8,1\}$
4	cat	$\{n_5,1\},\{n_6,1\},\{n_9,1\}$

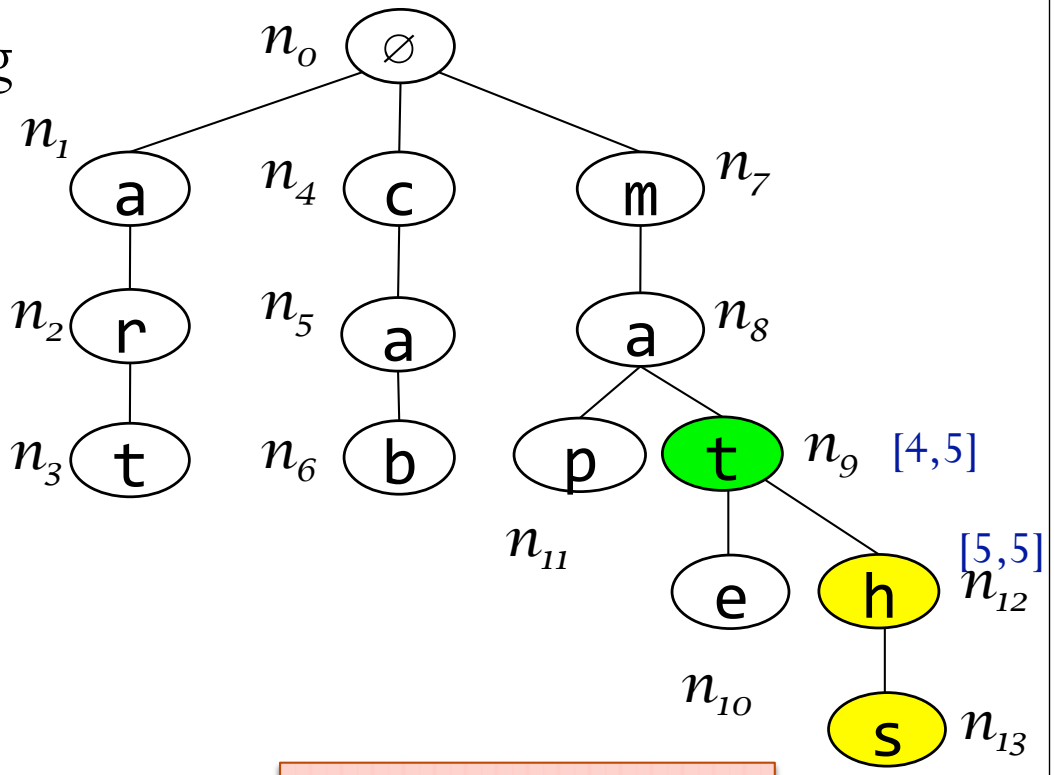
Step	Query	Active States and Their Extents
1	∅	$S_0 : \{n_0\}$
2	c	$S_0 : \{n_0\}$
3	ca	$S_2 : \{n_1\}, S_1 : \{n_4\}, S_5 : \{n_7\}$
4	cat	$S_1 : \{n_5\}, S_6 : \{n_8\}$



Q = cat, t = 1

Adaptation to Ed Search

- To ed search
  - DFSinstead of BFS
  - Result fetching: only retrieve leaf nodes
  - Extended length filtering
- To ed join
  - More involved



$EV(n_9) = S_7$ (aka. $[\#, \#, 1]^T$)
$ELenFilter(S_7, 1) = S_\perp$

prune n_9

Q = cat, t = 1

Parallelization

- Few published results AFAIK
- A poor man's approach
  - Ed-search:
    - partition the queries into fixed-size job block; each worker gets the next job block
  - Ed-join:
    - treated as batch edit similarity search

Conclusions

- Ed search/join is a HARD problem, yet still have very efficient methods for many practical settings
- Our preliminary study of trie-based methods for edit similarity queries
  - Small index size and pretty fast query processing speed for short string collections

- Lessons learned
  - Many open problems identified for (our) trie-based approach (e.g., long strings? large t?)
  - No one-size-fits-all solution (e.g., $|\Sigma|$ size, distributions)
  - Implementation details matter (e.g., parameter tuning?)

Q & A

More info @ our project Homepage:
http://www.cse.unsw.edu.au/~weiw/project/simjoin.html

Advertisement: ICDE2014 "Strings, Texts and Keyword Search" track