Effective and Scalable Solutions for Mixed and Split Citation Problems in Digital Libraries

Dongwon Lee, Byung-Won On
Penn State University, USA

Jaewoo Kang
North Carolina State University, USA

Sanghyun Park
Yonsei University, KOREA

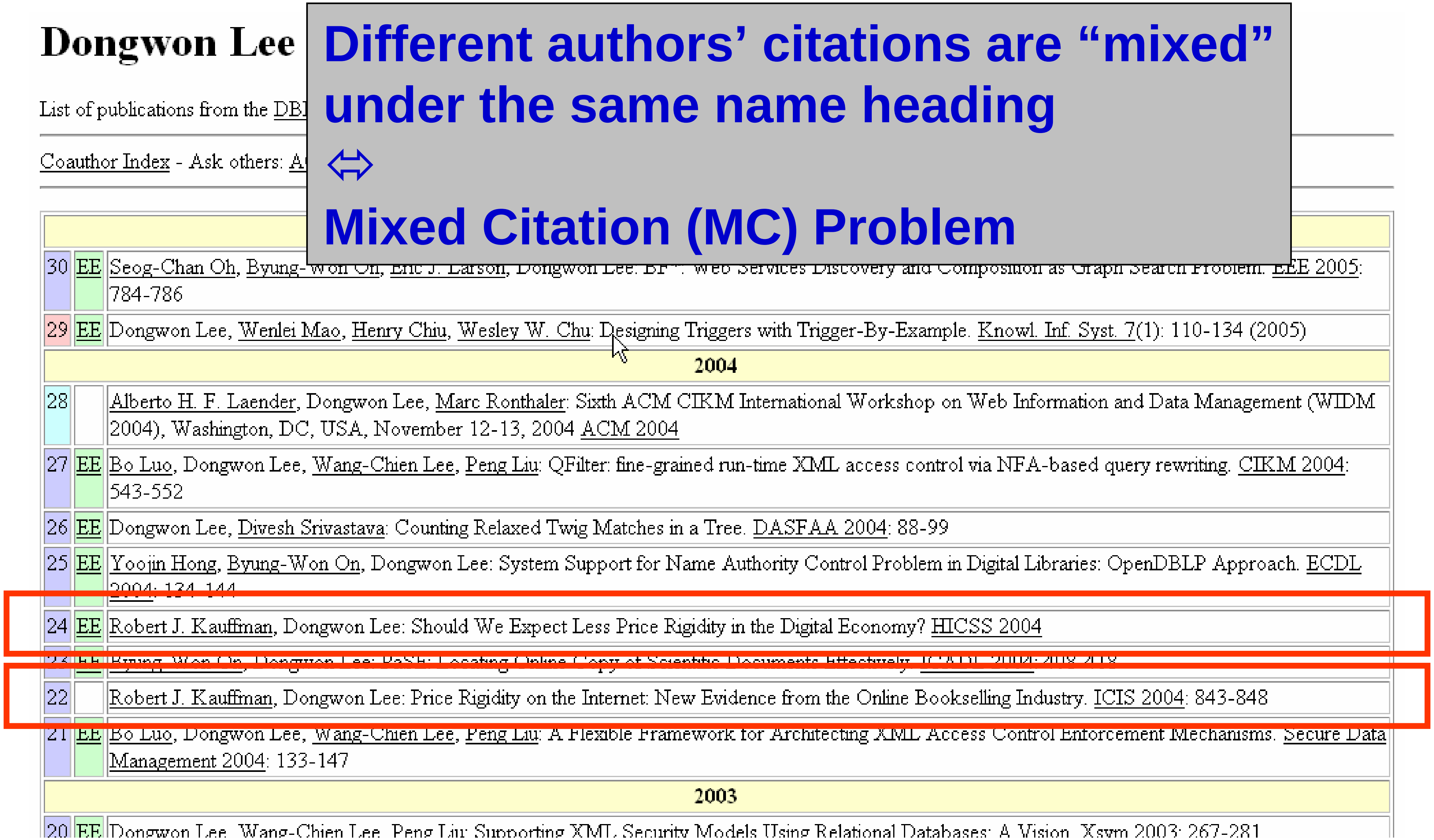
Outline

- Motivation
- Mixed Citation (MC) Problem
- Split Citation (SC) Problem
  - Problem Definition
  - Our approach
  - Preliminary Experimentation
- Summary

Motivation

- Digital Libraries (DL) often have many errors that negatively affect:
  - Quality of DL
  - Query results
  - User experiences
  - Bibliometric research
  - …
- We present 2 specific problems that often occur in *scientific literature* DL

Eg. 1: DBLP



Dongwon Lee

List of publications from the DBL

Coauthor Index - Ask others: A

30	EE	Seog-Chan Oh, Byung-Won On, Eric J. Larson, Dongwon Lee: BF*: Web Services Discovery and Composition as Graph Search Problem. EEE 2005: 784-786
29	EE	Dongwon Lee, Wenlei Mao, Henry Chiu, Wesley W. Chu: Designing Triggers with Trigger-By-Example. Knowl. Inf. Syst. 7(1): 110-134 (2005)
		2004
28		Alberto H. F. Laender, Dongwon Lee, Marc Ronthaler: Sixth ACM CIKM International Workshop on Web Information and Data Management (WIDM 2004), Washington, DC, USA, November 12-13, 2004 ACM 2004
27	EE	Bo Luo, Dongwon Lee, Wang-Chien Lee, Peng Liu: QFilter: fine-grained run-time XML access control via NFA-based query rewriting. CIKM 2004: 543-552
26	EE	Dongwon Lee, Divesh Srivastava: Counting Relaxed Twig Matches in a Tree. DASFAA 2004: 88-99
25	EE	Yoojin Hong, Byung-Won On, Dongwon Lee: System Support for Name Authority Control Problem in Digital Libraries: OpenDBLP Approach. ECDL 2004: 134-144
24	EE	Robert J. Kauffman, Dongwon Lee: Should We Expect Less Price Rigidity in the Digital Economy? HICSS 2004
23	EE	Byung-Won On, Dongwon Lee: PaSE: Locating Online Copy of Scientific Documents Effectively. ICADL 2004: 408-418
22		Robert J. Kauffman, Dongwon Lee: Price Rigidity on the Internet: New Evidence from the Online Bookselling Industry. ICIS 2004: 843-848
21	EE	Bo Luo, Dongwon Lee, Wang-Chien Lee, Peng Liu: A Flexible Framework for Architecting XML Access Control Enforcement Mechanisms. Secure Data Management 2004: 133-147
		2003
20	EE	Dongwon Lee, Wang-Chien Lee, Peng Liu: Supporting XML Security Models Using Relational Databases: A Vision. Xsym 2003: 267-281

Different authors' citations are "mixed" under the same name heading

⇔

Mixed Citation (MC) Problem

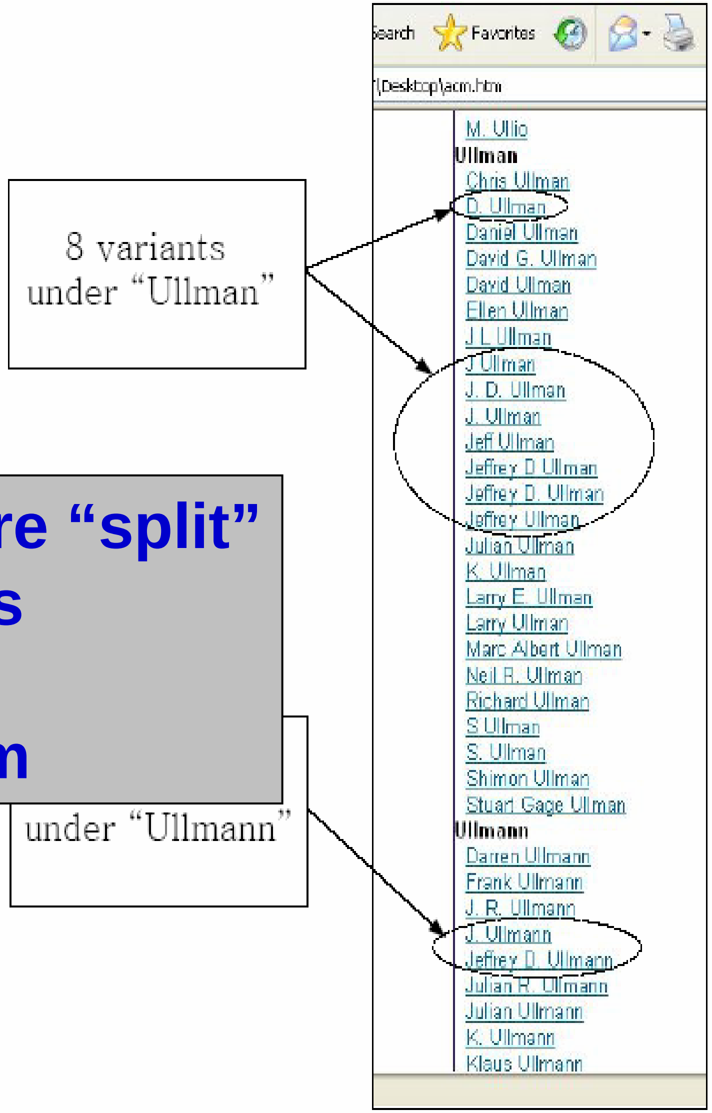
1. Mixed Citation Problem

- Given a collection of citations (*C*) by an author (*a_i*), can we identify false citations by another author (*a_j*), when *a_i* and *a_j* have the identical name spellings (i.e., homonym)?

- Solution: Citation Labeling Algorithm
- Idea: for each citation in the collection, test if the citation really belongs to the given collection

Eg. 2: ACM DL Portal

Jeffrey D. Ullman
@ Stanford Univ.



8 variants under "Ullman"

under "Ullmann"

M. Ullio
Ullman
Chris Ullman
D. Ullman
Daniel Ullman
David G. Ullman
David Ullman
Ellen Ullman
J L Ullman
J Ullman
J. D. Ullman
J. Ullman
Jeff Ullman
Jeffrey D Ullman
Jeffrey D. Ullman
Jeffrey Ullman
Julian Ullman
K. Ullman
Larry E. Ullman
Larry Ullman
Marc Albert Ullman
Neil R. Ullman
Richard Ullman
S Ullman
S. Ullman
Shimon Ullman
Stuart Gage Ullman
Ullmann
Darren Ullmann
Frank Ullmann
J. R. Ullmann
J. Ullmann
Jeffrey D. Ullmann
Julian R. Ullmann
Julian Ullmann
K. Ullmann
Klaus Ullmann

Same authors' citations are "split" into various name variants
⇔
Split Citation (SC) Problem

2. Split Citation (SC) Problem

- Given two lists of author names, X and Y, for each author name x ($\in X$), find a set of author names, y_1, y_2, …, y_n ($\in Y$) such that both x and y_i ($1 \le i \le n$) are variants

2. Split Citation (SC) Problem

"tuple"

- Given two lists of ~~author names,~~ X and Y, for each ~~author name~~ x ($\in X$), find a set of ~~author names,~~ ~~y_1, y_2,~~ …, y_n ($\in Y$) such that both x and y_i ($1 \leq i \leq n$) are variants

= Database Join Problem

2. Split Citation (SC) Problem

“record”

- Given two lists of ~~author names,~~ X and Y, for each ~~author name~~ x ($\in X$), find a set of author ~~names, y_1, y_2,~~ …, y_n ($\in Y$) such that both x and y_i ($1 \leq i \leq n$) are variants

= Record Linkage Problem

Naïve Solution

- For each author name *x* in *X*
  - For each author name *y* in *Y*

    If *x* ~ *y*, name variant !

$O(|X||Y|)$

dist(x,y) < t

- DB: Nested Loop Join
- RL: Pair-wise Record Match

Challenge 1: Scalability

- O(|X||Y|) is too costly
- Solutions
  - DB: Hashed Join
  - RL: Blocking

- For each name *x* in *X*
  - Assign *x* to block *b* (∈*B*)
- For each name *y* in *Y*
  - Assign y to block *b* (∈*B*)
- For each block *b* (∈*B*)
  - *Do naïve-solution*

DL	Domain	Size (in M)
ISI/SCI	General Sciences	25
CAS	Chemistry	23
Medline/ PubMed	Life Science	12
CiteSeer	General Sciences/ Engineering	10
arXiv	Physics/Math	0.3
SPIRED HEP	Physics	0.5
DBLP	CompSci	0.6
CSB	CompSci	1.4

$O(|X|+|Y|+|B|a) << O(|X||Y|)$

Challenge 2: Distance

- Diverse name variations
  - "Jeffrey D. Ullman" ⇔ "J. Ullman"
  - "Alon Y. Levy" ⇔ "Halevy, A."
  - "W. Wang" ⇔ "X. Wang"
  - "Sean Engelson" ⇔ "Shlomo Argamon"
- Solution
  - Look at additional information of the author names
  - Eg,

    Coauthor list

    Keywords used in title

    Venues to submit

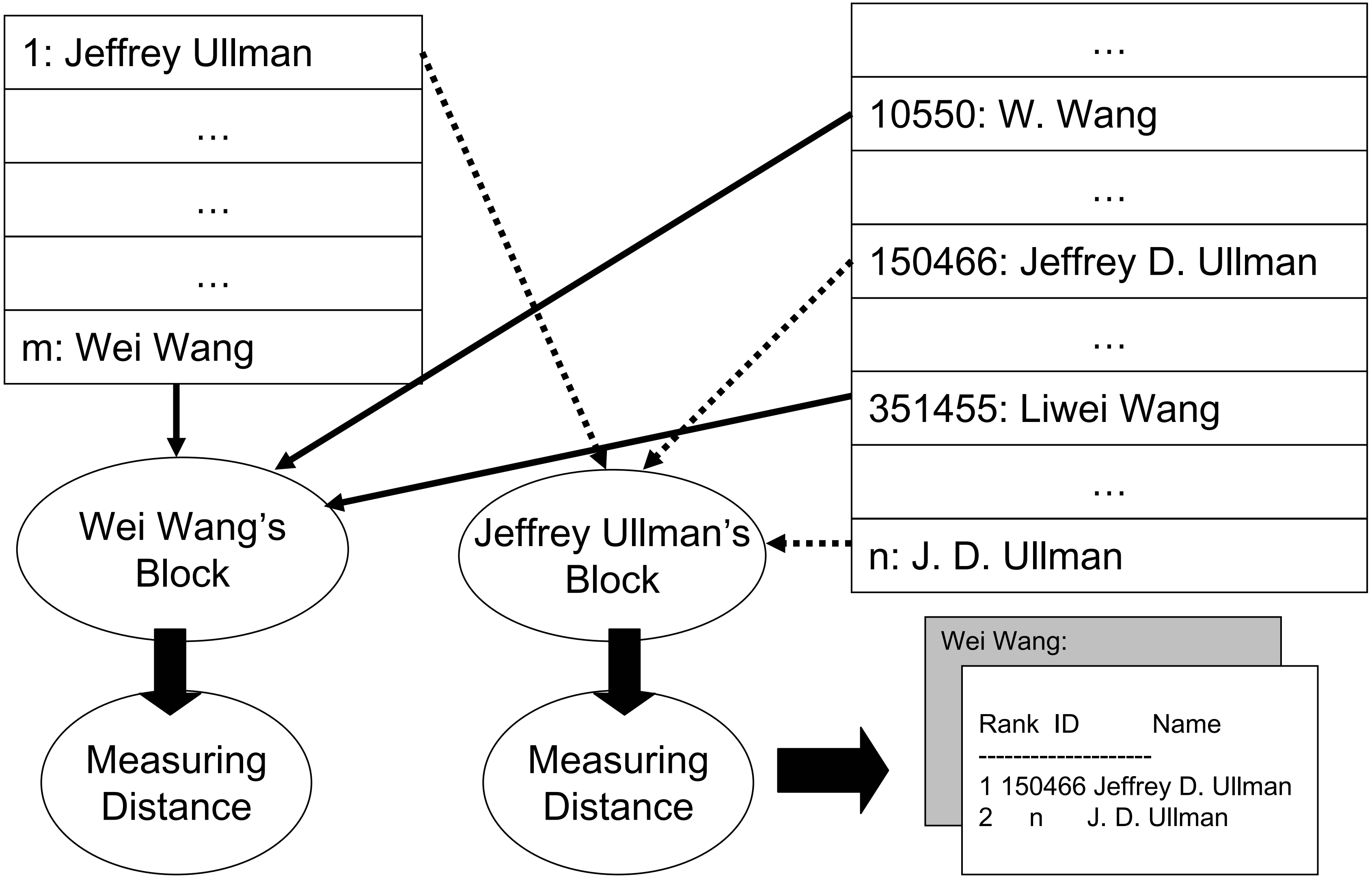
    Year

    Affiliation

    …

$$
\begin{aligned}
&dist(x,y) \sim \\
&\qquad W_i * dist(C(x),C(y)) + \\
&\qquad W_j * dist(T(x),T(y)) + \\
&\qquad W_k * dist(V(x),V(y))
\end{aligned}
$$

Name Disambiguation Algorithm



1: Jeffrey Ullman

…

…

…

m: Wei Wang

…

10550: W. Wang

…

150466: Jeffrey D. Ullman

…

351455: Liwei Wang

…

n: J. D. Ullman

Wei Wang:

```
Rank  ID          Name
--------------------
1 150466 Jeffrey D. Ullman
2     n       J. D. Ullman
```

Step 1: Blocking

- Many blocking methods can be applied
  - Sorted Window
  - Token-based
  - N-gram
  - Sampling
- We applied Gravano (2003)'s sampling-based join approximation algorithm as a blocking method
  - Comparison with other blocking methods => JCDL 2005

Step 2: Measuring Distance

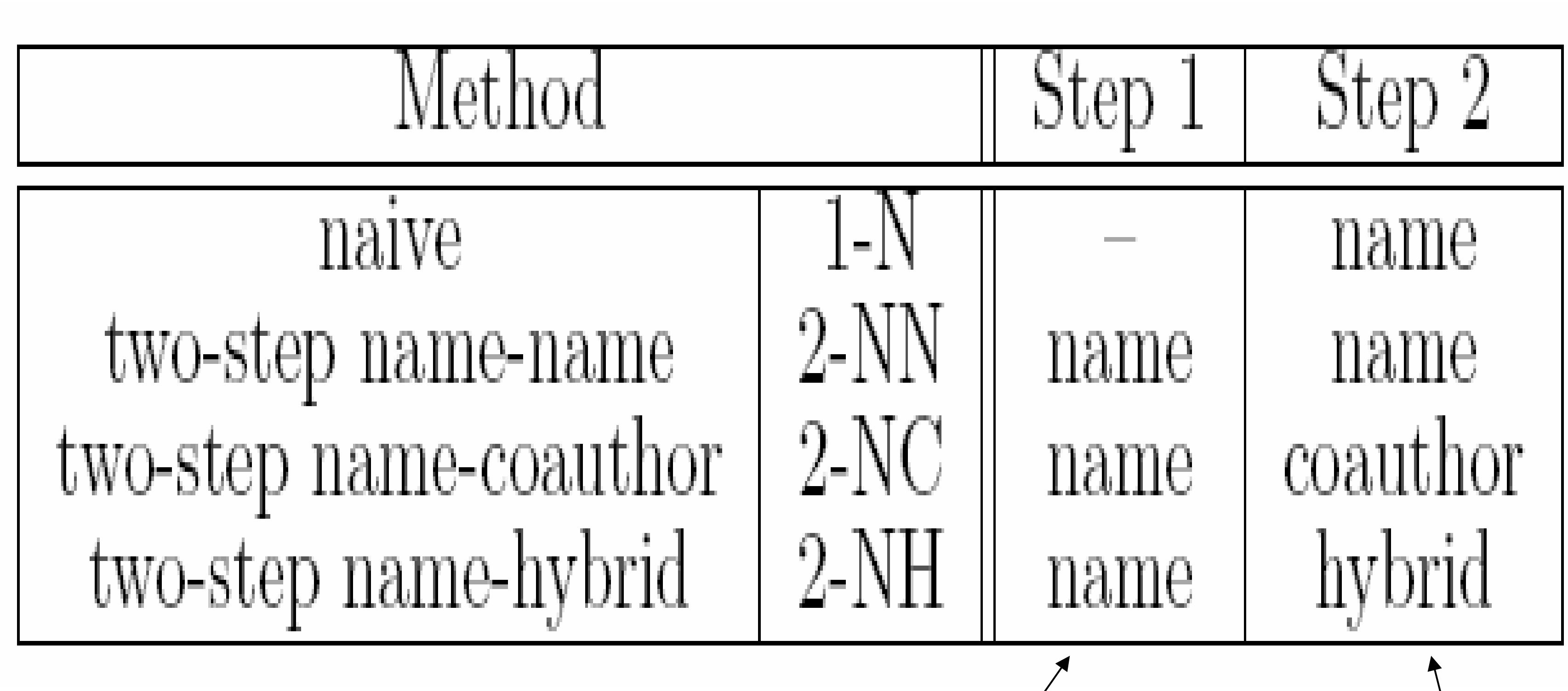
- Naïve Bayes Model
  - Use Bayes’ Theorem to measure similarity between two names
- Support Vector Machine
  - Use SVM Classifiers

Supervised

- String-based Distance Metrics
  - TFIDF/Jaccard (Token-based)
  - Jaro/JaroWinkler (Edit distances)
- Vector-based Cosine Distance
  - Cosine Similarity

Un-supervised

Policy Variations

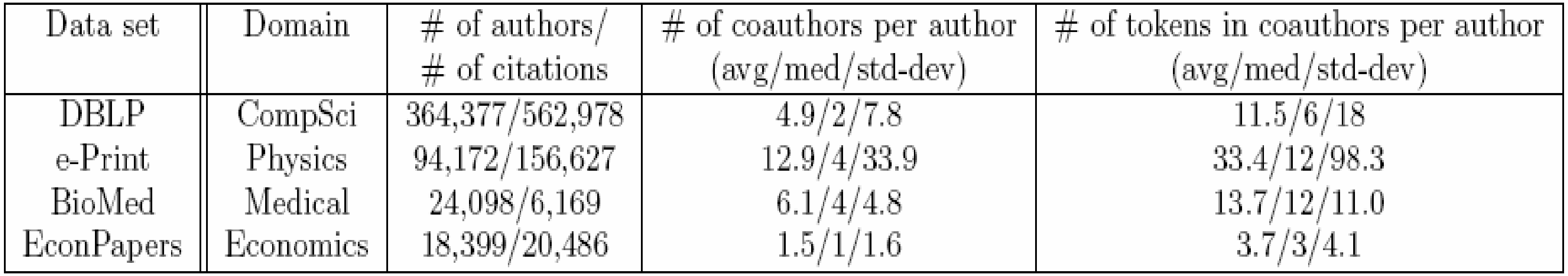


Method		Step 1	Step 2
naive	1-N	–	name
two-step name-name	2-NN	name	name
two-step name-coauthor	2-NC	name	coauthor
two-step name-hybrid	2-NH	name	hybrid

Blocking

Measuring Distance

Data sets



Data set	Domain	# of authors/ # of citations	# of coauthors per author (avg/med/std-dev)	# of tokens in coauthors per author (avg/med/std-dev)
DBLP	CompSci	364,377/562,978	4.9/2/7.8	11.5/6/18
e-Print	Physics	94,172/156,627	12.9/4/33.9	33.4/12/98.3
BioMed	Medical	24,098/6,169	6.1/4/4.8	13.7/12/11.0
EconPapers	Economics	18,399/20,486	1.5/1/1.6	3.7/3/4.1

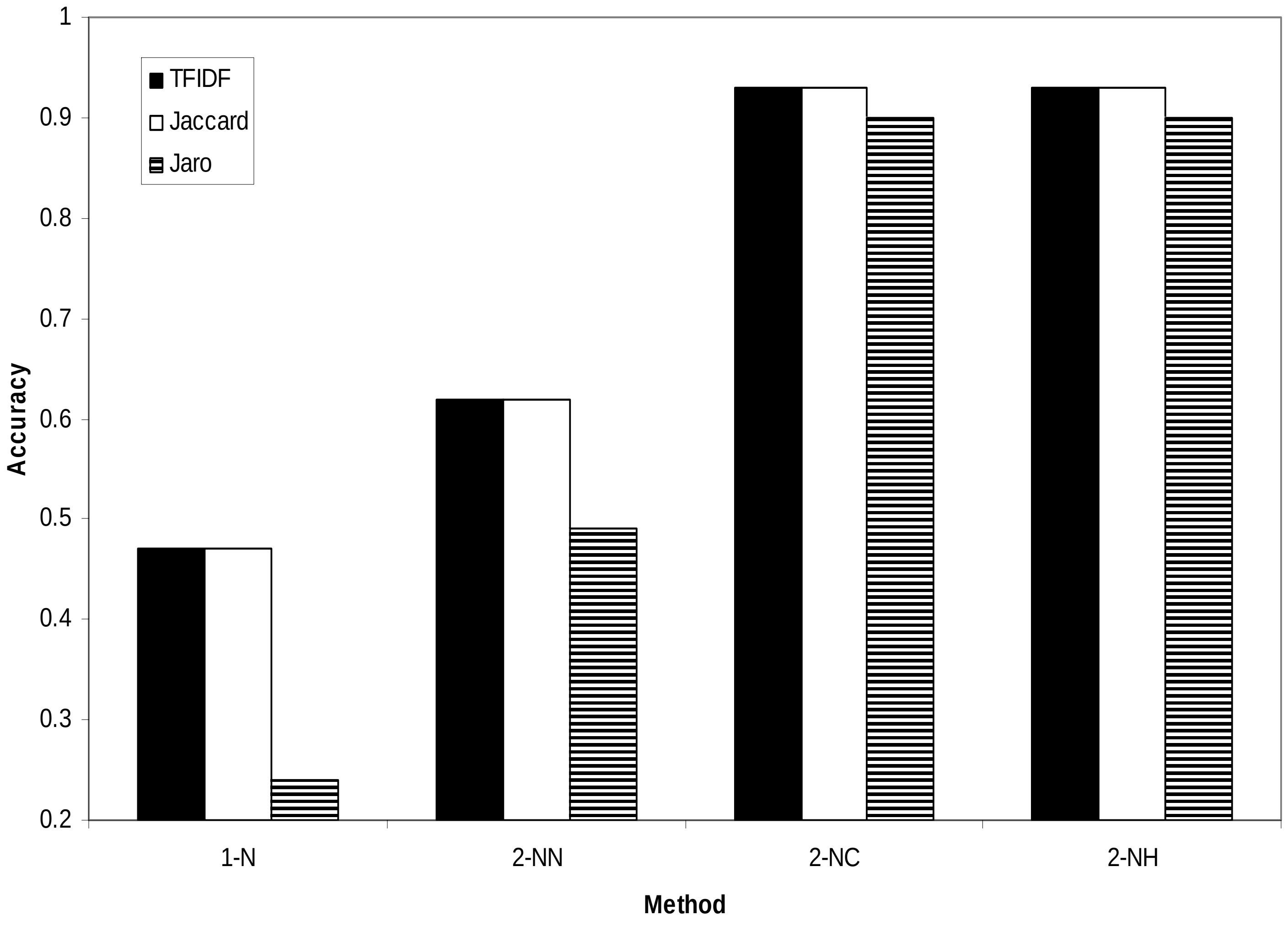
Configuration (eg, DBLP case)

- Authors, x, in *X* and authors, y, in *Y*
- Prepare an artificial name variant x’ for K randomly-chosen x (eg, K=100):
  - Abbreviation of the first name (85%): “Ji-Woo K. Li” → “J. K. Li”
  - Typo (15%): “Ji-Woo K. Li” → “Ji-Woo K. Lee”
  - x’ carries half of x’s original citations
  - x carries the other half
  - Inject all x’ into Y
- Test*: “for each author x in X, fi*
  name variants x’ in Y”
- Evaluation metrics
  - Time
  - Accuracy

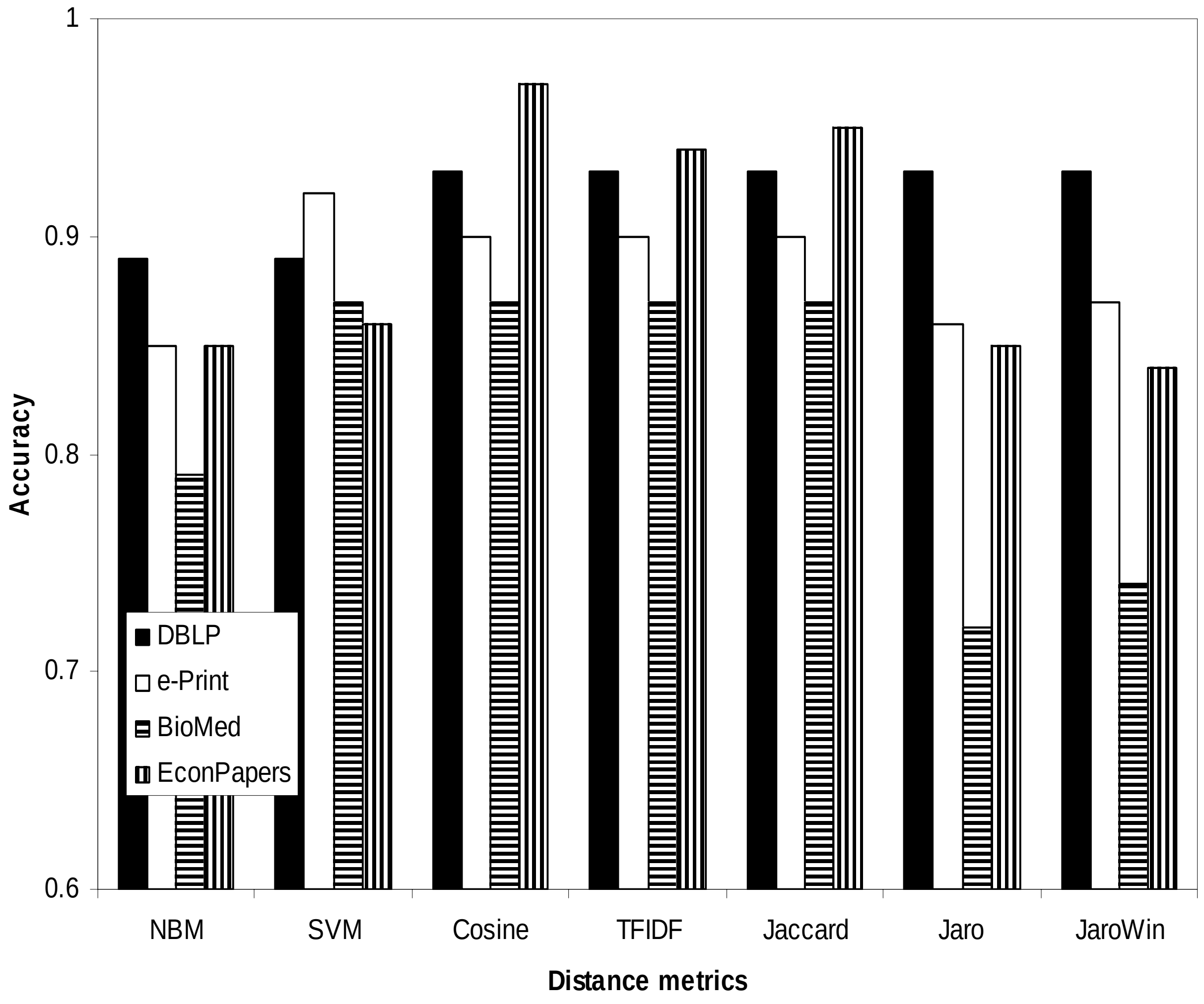
Varying error types gave consistent results. For instance,

Name Abbreviation: 30%
Name Alternation: 30%
First Name Misspelling: 12%
Last Name Misspelling: 12%
Contraction: 2%
Middle Name Initial Omission: 4%
Combination: 10%

SC: Accuracy (DBLP)



SC: Accuracy (All data sets)



Related Work

- Identity / Entity Matching
  - Database Join
  - Record Linkage
  - Merge / Purge
  - Ontology Matching
  - Graph Matching
  - …
- Name Authority Control Problem in LIS
- Please see the paper for details

Future Work

- Using additional information of author name
  - Essentially, token comparison
- Better way: coauthor information as a Graph
  - Graph matching / partitioning
  - Sub-graph detection

Conclusion

- SC Problem
  - Using additional information (eg, coauthor) than name itself is better in distance measure
  - 2-NC/2-NH outperform 1-N/2-NN
  - *SVM or Cosine* shows the best accuracy (90-93%)