

System Support for Name Authority Control Problem in Digital Libraries: OpenDBLP Approach

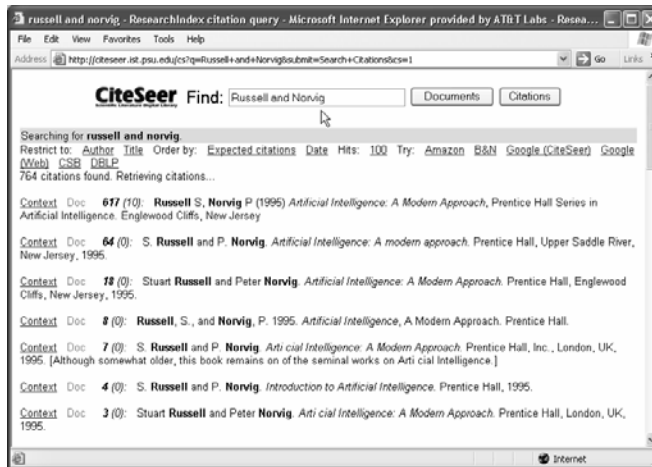
Yoojin Hong
Byung-Won On
Dongwon Lee

Penn State University, USA



Motivation I

PENNSTATE





Motivation I

- Search result of CiteSeer for “**Artificial Intelligence: A Modern Approach**” by *Russell and Norvig*
- CiteSeer lists 24 citations as different
- But all of them are variations, and must be consolidated



Motivation I

- This is well-known “Citation Matching” problem
- Also known as
 - Record Linkage
 - Identity Uncertainty
 - Merge/Purge
 - Object De-duplication
 - Database Join
 - ...



Motivation I

- The problem occurs mainly due to
 - Different formats that people use
 - Erroneous data entry
 - Imperfect citation gathering software
 - ...
- These are **Syntactic** variations



Motivation II

- *Alon Levy*, U. Washington, got married and changed his last name to *Alon Halevy*
- Two *Dongwon Lee*
 - Penn State (Database), U. Minnesota (MIS)
- ACM *DL* and IEEE *ADL* merged into ACM/IEEE *JCDL* in 2001
- *ACL* and *COLING* merged into *ACL-COLING* in 1998 and split afterwards
- Known as “Name Authority Control”



Motivation II

- One way or the other, all are **Name** related
- All changes are legitimate, but unavoidable as time passes
- These are **Semantic** variations

- Digital Libraries must keep track of these changes to provide better services to users
 - Eg, search result show related names

I. Existing Solutions for **Syntactic** Variations



- Periodically run algorithms to detect syntactic variations
 - Record linkage or citation matching algorithm
 - Some claim to achieve 80-95% accuracy for limited domain
- Actively being researched by large community
- But not much research as to “**What Next?**”

II. Existing Solutions for Semantic Variations



- Libraries keep variations in authority file
- When needed, consult the file to disambiguate variations

- Not much systematic help for users

Issues of Existing Solutions



- By an large, focused on the **Identification** of syntactic or semantic name irregularities
 - DOI, OpenURL, ISBN, ...
 - Open Journal Project, Open Citation Project, ...

- Not much efforts on **how systems can utilize** the learned knowledge on name variations



Problem Definition

*In Digital Libraries, when bibliographic entities (e.g., author, publication venue, etc) evolve over time, **given the changes are known**, devise a system support such that systems can **update** and **search** the changes easily.*



Significance

- Large number
 - NSF DL Initiative, Europe DELOS
- Large size
 - CiteSeer: 10 M CAS: 23 M
 - ISI/SCI: 25 M PubMed: 10 M ...
- Automatically constructed – tend to have more citation errors and name variations
 - CiteSeer, eBizSearch, CSB, BibFinder, ...



Contributions

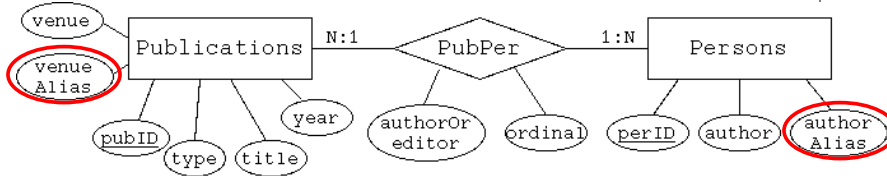
- Identification of three core building blocks for name change patterns
- Taxonomy of their combinations
- Simple system architecture to support UPDATE and SEARCH
- Proof of concept tested in OpenDBLP



UPDATE: Building Blocks

- We view all name authority controls as three patterns:
 - Linear change: $A \Rightarrow B$
 - Split: $A \Rightarrow \{A1, A2\}$
 - Merge: $\{A1, A2\} \Rightarrow A$

UPDATE: Database Support



1. (100, "Levy", null) =>
 - (100, "Levy", "Halevy")
 - (101, "Halevy", null)
2. (200, "Dongwon Lee", null) =>
 - (200, "Dongwon Lee", null)
 - (201, "Dongwon Lee", null)
3. (300, "Corator", null), (301, "D. Corator", null) =>
 - (300, "Corator", "Lee D. Corator")
 - (301, "D. Corator", "Lee D. Corator")
 - (302, "Lee D. Corator", null)

SEARCH: System Support

- In searching answers matching users' queries, systems must exploit the knowledge of name variations
- Eg, "Retrieve all publications about XML by Alon Halevy"



SEARCH: System Support

```
(SELECT P1.*
FROM Publications P1, PubPer P2, Persons P3
WHERE P1.pubID=P2.pubID AND P2.perID=P3.perID AND
P3.author = 'Alon Halevy' and
P1.title LIKE '%XML%')
```



SEARCH: System Support

```
(SELECT P1.*
FROM Publications P1, PubPer P2, Persons P3
WHERE P1.pubID=P2.pubID AND P2.perID=P3.perID AND
P3.author = 'Alon Halevy' and
P1.title LIKE '%XML%')

UNION

(SELECT P1.*
FROM Publications P1, PubPer P2, Persons P3,
Persons P4
WHERE P1.pubID=P2.pubID AND P2.perID=P4.perID AND
P3.authorAlias = 'Alon Halevy' AND
P1.title LIKE '%XML%' AND
P3.authorAlias = P4.author)
```



SEARCH: System Support

- Different choices on schema
 - Alias as separate table
 - Alias as ID, not as values
 - ...
- SQL for SEARCH would be different then
- SEARCH can be easily extended to exploit the name variations information, if those are UPDATED properly.

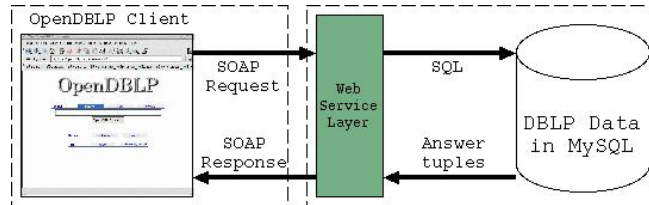


SEARCH: Strategy

- Three temporal strategies: Backward, Forward, Semantic
 - $C1 \Rightarrow C2$
 - $C3 \Rightarrow \{C4, C5\}$
 - $\{C6, C7\} \Rightarrow C8 \Rightarrow C9$
- **Backward**:
 - User searches for C2
 - System returns C2 and all its predecessors, C1
- **Forward**: symmetric case of Backward
- **Semantic** = Backward + Forward
 - User searches for C8
 - Systems returns $\{C6, C7\} + \{C9\}$



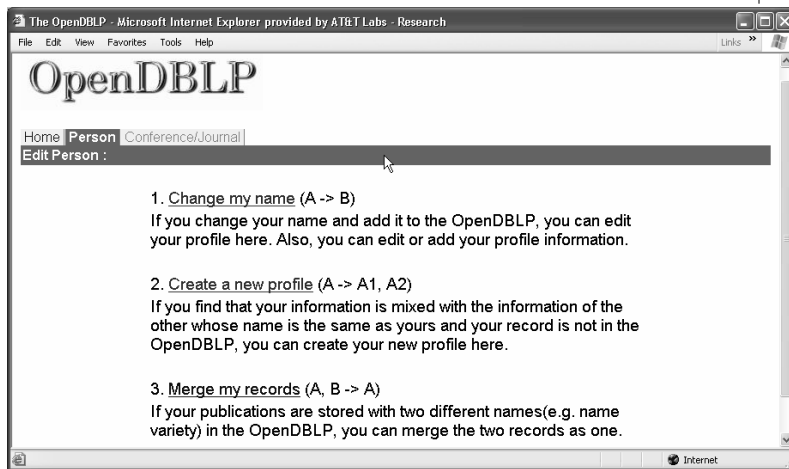
Implementation: OpenDBLP




- UPDATE and SEARCH functions are implemented in the prototype system, OpenDBLP
- OpenDBLP: Based on DBLP digital library
 - Small: 0.5 M
 - CompSci domain only



Demonstration: Menu



Demonstration: Linear Change



OpenDBLP

Home | **Person** | Conference/Journal

Edit Person : 'Alon Y. Halevy'

Personal Information

Name

Alias

Homepage

Org

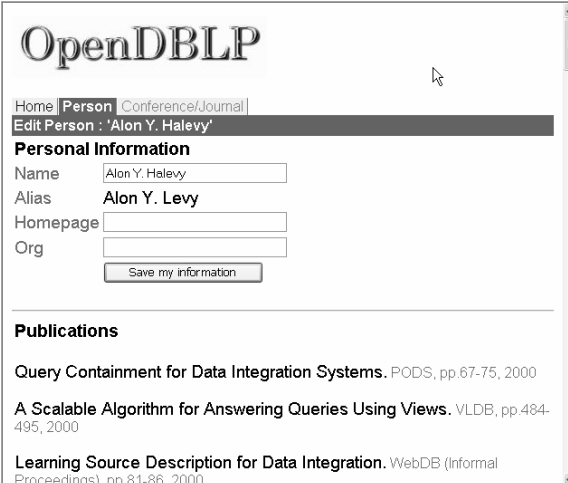
Publications

Composing Mappings Among Data Sources. VLDB, pp.572-583, 2003

Query containment for data integration systems. JCSS, volume 66, 2003

Semantic Email: Adding Lightweight Data Manipulation Capabilities to

Demonstration: Linear Change



OpenDBLP

Home | **Person** | Conference/Journal

Edit Person : 'Alon Y. Halevy'

Personal Information

Name

Alias

Homepage

Org

Publications

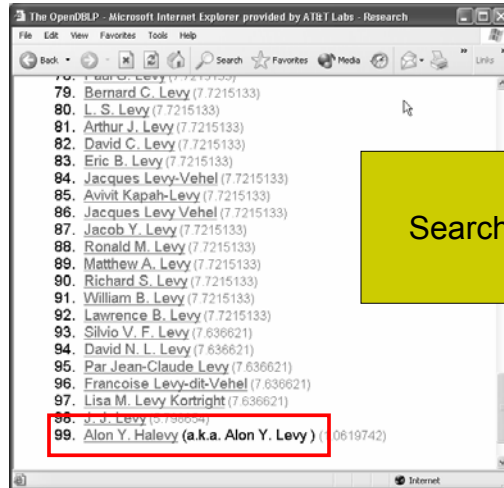
Query Containment for Data Integration Systems. PODS, pp.67-75, 2000

A Scalable Algorithm for Answering Queries Using Views. VLDB, pp.484-495, 2000

Learning Source Description for Data Integration. WebDB (Informal Proceedings), pp.81-86, 2000



Demonstration: Linear Change



Search: "Levy"



Demonstration: Split





Demonstration: Split

OpenDBLP

Home | **Person** | Conference/Journal

Edit Person : Creating a new profile for 'Wei Wang'

Personal Information - Input your information

Name Wei Wang

Homepage

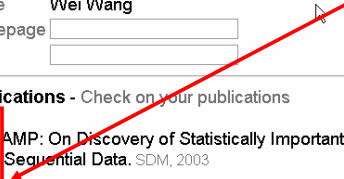
Org

Publications - Check on your publications

- STAMP: On Discovery of Statistically Important Pattern Repeats in Long Sequential Data. SDM, 2003
- Enhanced Biclustering on Expression Data. BIBE, pp.321-327, 2003
- ApproxMAP: Approximate Mining of Consensus Sequential Patterns. SDM, 2003

ECDL 2004 27

Select all articles
By "Wei Wang"
At HKUST



Demonstration: Split



OpenDBLP

Home | **Person** | Conference/Journal

Edit Person : Creating a new profile for 'Wei Wang'

Your new profile is successfully created!

Personal Information

Personal ID 288721

Name Wei Wang

Alias

Homepage

org

Publications

ApproxMAP: Approximate Mining of Consensus Sequential Patterns. SDM, 2003

STAMP: On Discovery of Statistically Important Pattern Repeats in

ECDL 2004 28

New record for
"Wei Wang"
at KHUST

Demonstration: Split



OpenDBLP

wei wang
OpenDBLP Search

Home Person Title Keywords

Search results for 'wei wang': All 100 results. Search took 1.492 seconds.

Authors/Editors that match the search string:

1. Wei Wang (11.288343)
2. Wei Wang (11.288343)
3. Wang wei (11.288343)
4. Wei-lung Wang (11.162856)
5. Jian-Wei Wang (11.162856)

Demonstration

<http://opendbpl.psu.edu/>



Future Work

- Further testing on different domain and size
 - arXiv e-Print
 - CiteSeer
- Efficient implementation of SEARCH
- The knowledge of semantic variations must be exploited in identifying semantic variations
 - Better citation matching algorithms
- Application of Query Expansion technique
 - User profile, previous search history, ...



Conclusion

Digital Libraries must exploit **identified** and **updated** name variations to provide better services.

Thanks !!