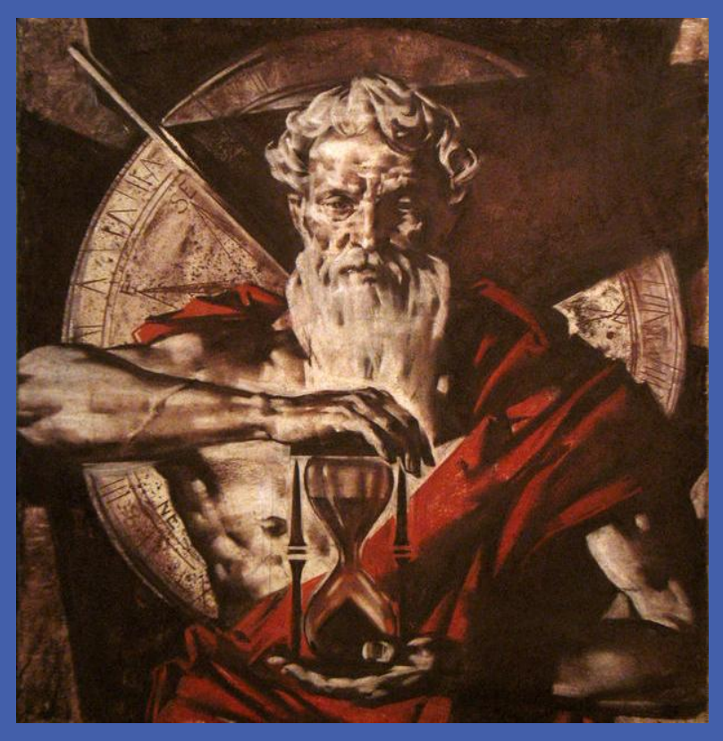


CHRONOS: **Facilitating History Discovery by Linking Temporal Records**

Pei Li[1], Haidong Wang[2], Christina Tziviskou[1], Xin Luna Dong[3], Xiaoguang Liu[2], Andrea Maurino[1], Divesh Srivastava[3]

University of Milan – Bicocca[1], Nankai University[2], AT&T LabsResearch[3]

Motivations

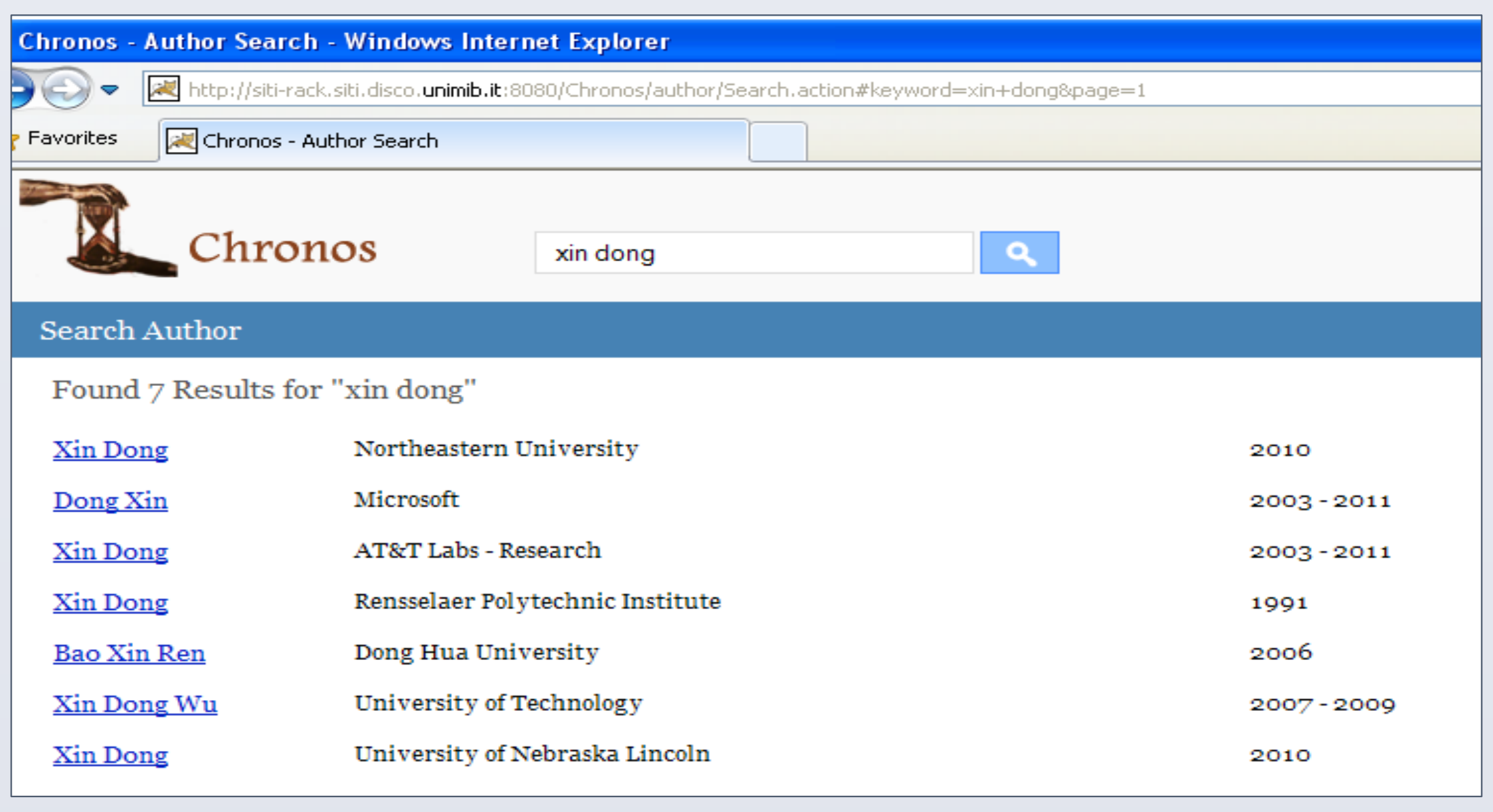
Many data sets contain temporal records over a long period of time; each record is associated with a time stamp and describes some aspects of a real-world entity at that particular time. From such data, users often wish to search for entities in a particular period and understand the history of one entity or all entities in the data set. A major challenge for enabling such search and exploration is to identify records that describe the same real-world entity over a long period of time; however, linking temporal records is hard given that the values that describe an entity can evolve over time (e.g., a person can move from one affiliation to another).

We demonstrate the CHRONOS system which offers users the useful tool for finding real-world entities over time and understanding history of entities in the bibliography domain. The core of CHRONOS is a temporal record-linkage algorithm [1], which is tolerant to value evolution over time. CHRONOS (1) allows users to explore the history of authors, (2) helps users understand linkage results by comparing our results with those of existing systems, highlighting differences in the results, explaining our decisions to users, (3) and answering "what-if" questions.

System Features

Searching author:

CHRONOS supports keyword search on author name, affiliation and publishing year. The snapshot shows the results of searching "Xin Dong": it shows 7 authors, each with name, current affiliation and publishing period.



Chronos - Author Search - Windows Internet Explorer
http://siti-rack.siti.disco.unimib.it:8080/Chronos/author/Search.action#keyword=xin+dong&page=1
Favorites
Chronos - Author Search
Chronos
xin dong

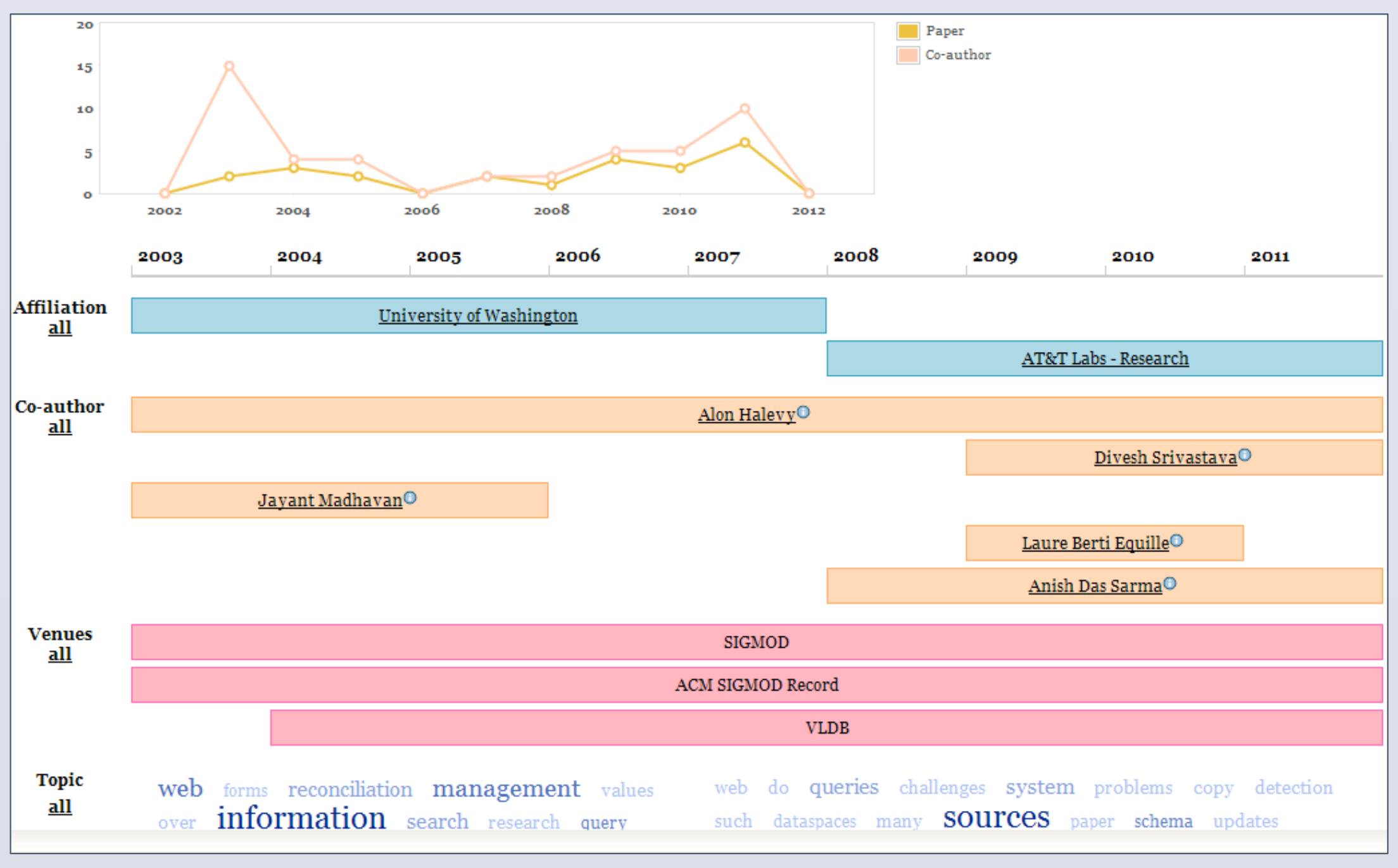
Search Author

Found 7 Results for "xin dong"

Xin Dong	Northeastern University	2010
Dong Xin	Microsoft	2003 - 2011
Xin Dong	AT&T Labs - Research	2003 - 2011
Xin Dong	Rensselaer Polytechnic Institute	1991
Bao Xin Ren	Dong Hua University	2006
Xin Dong Wu	University of Technology	2007 - 2009
Xin Dong	University of Nebraska Lincoln	2010

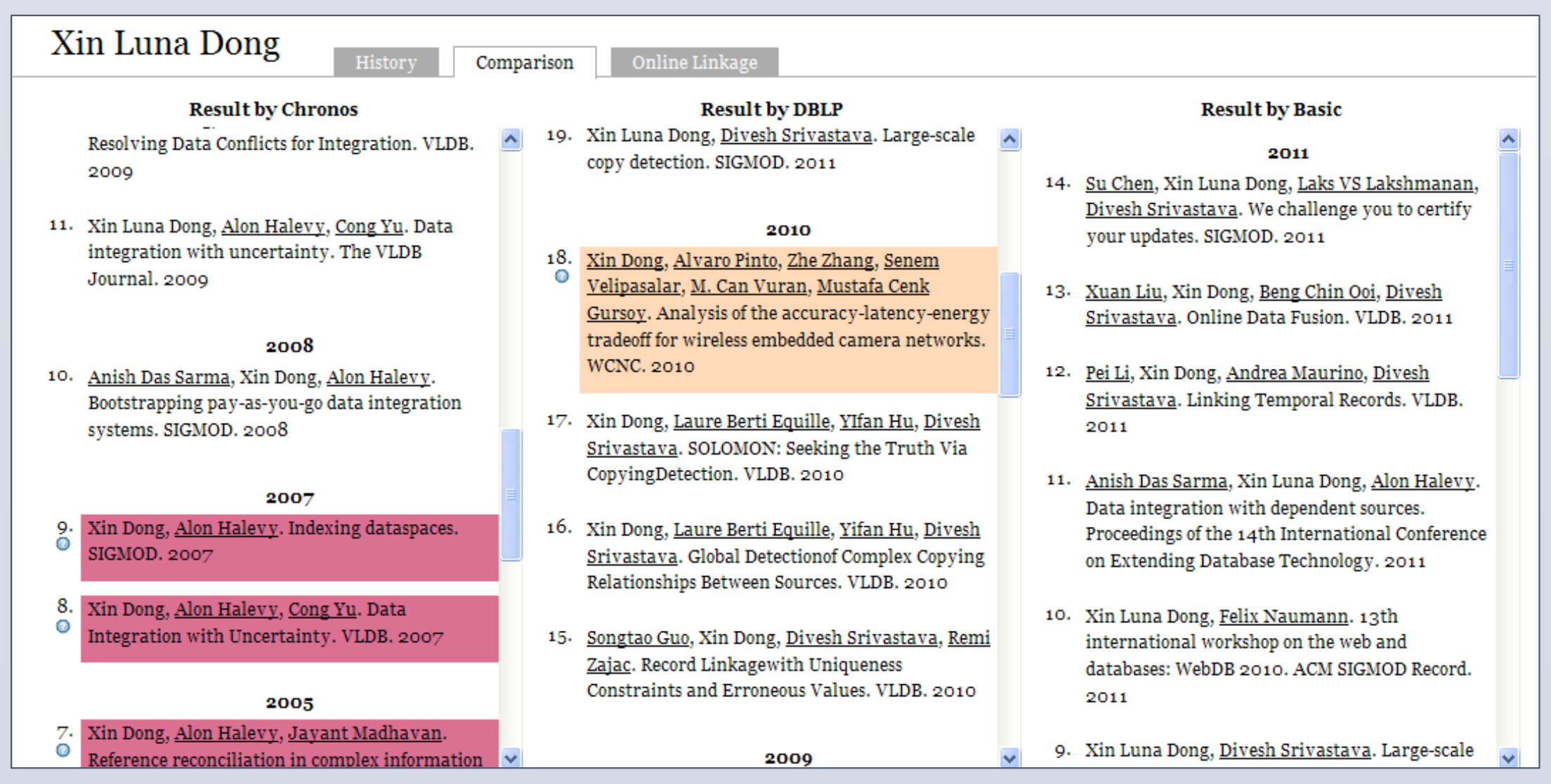
Tracing history:

CHRONOS allows to trace the history of a particular author, such as her affiliation, co-authors, research topics, and so on. The snapshot shows the history of various aspects of author Xin Dong, and her publishing statistics .



Comparing results:

For each author, CHRONOS shows side-by-side the list of papers according to the linkage results by CHRONOS, by DBLP, and by BASIC [2]. It also highlights differences between the lists.



Result by Chronos

Resolving Data Conflicts for Integration. VLDB. 2009

11. Xin Luna Dong, Alon Halevy, Cong Yu. Data integration with uncertainty. The VLDB Journal. 2009

2008

10. Anish Das Sarma, Xin Dong, Alon Halevy. Bootstrapping pay-as-you-go data integration systems. SIGMOD. 2008

2007

9. Xin Dong, Alon Halevy. Indexing dataspaces. SIGMOD. 2007
8. Xin Dong, Alon Halevy, Cong Yu. Data Integration with Uncertainty. VLDB. 2007

2005

7. Xin Dong, Alon Halevy, Jayant Madhavan. Reference reconciliation in complex information

Result by DBLP

19. Xin Luna Dong, Divesh Srivastava. Large-scale copy detection. SIGMOD. 2011

2010

18. Xin Dong, Alvaro Pinto, Zhe Zhang, Senem Velipasalar, M. Can Vuran, Mustafa Cenk Gursoy. Analysis of the accuracy-latency-energy tradeoff for wireless embedded camera networks. WCNC. 2010
17. Xin Dong, Laure Berti Equille, YIfan Hu, Divesh Srivastava. SOLOMON: Seeking the Truth Via CopyingDetection. VLDB. 2010
16. Xin Dong, Laure Berti Equille, Yifan Hu, Divesh Srivastava. Global Detectionof Complex Copying Relationships Between Sources. VLDB. 2010
15. Songtao Guo, Xin Dong, Divesh Srivastava, Remi Zajac. Record Linkagewith Uniqueness Constraints and Erroneous Values. VLDB. 2010

2009

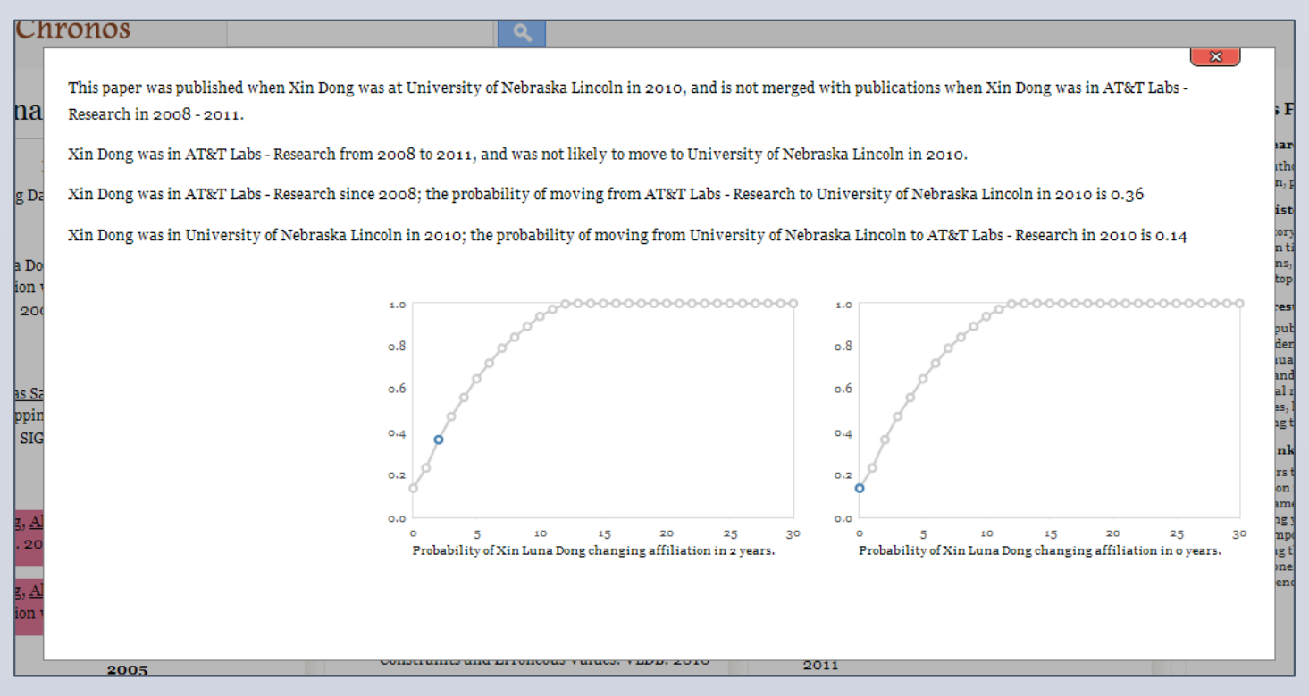
Result by Basic

2011

14. Su Chen, Xin Luna Dong, Laks VS Lakshmanan, Divesh Srivastava. We challenge you to certify your updates. SIGMOD. 2011
13. Xuan Liu, Xin Dong, Beng Chin Ooi, Divesh Srivastava. Online Data Fusion. VLDB. 2011
12. Pei Li, Xin Dong, Andrea Maurino, Divesh Srivastava. Linking Temporal Records. VLDB. 2011
11. Anish Das Sarma, Xin Luna Dong, Alon Halevy. Data integration with dependent sources. Proceedings of the 14th International Conference on Extending Database Technology. 2011
10. Xin Luna Dong, Felix Naumann. 13th international workshop on the web and databases: WebDB 2010. ACM SIGMOD Record. 2011
9. Xin Luna Dong, Divesh Srivastava. Large-scale

Explaining difference:

CHRONOS explains difference decisions on each highlighted publications. The snapshot explains why paper # 18 of Xin Luna Dong is included by DBLP but not by CHRONOS.



This paper was published when Xin Dong was at University of Nebraska Lincoln in 2010, and is not merged with publications when Xin Dong was in AT&T Labs - Research in 2008 - 2011.

Xin Dong was in AT&T Labs - Research from 2008 to 2011, and was not likely to move to University of Nebraska Lincoln in 2010.

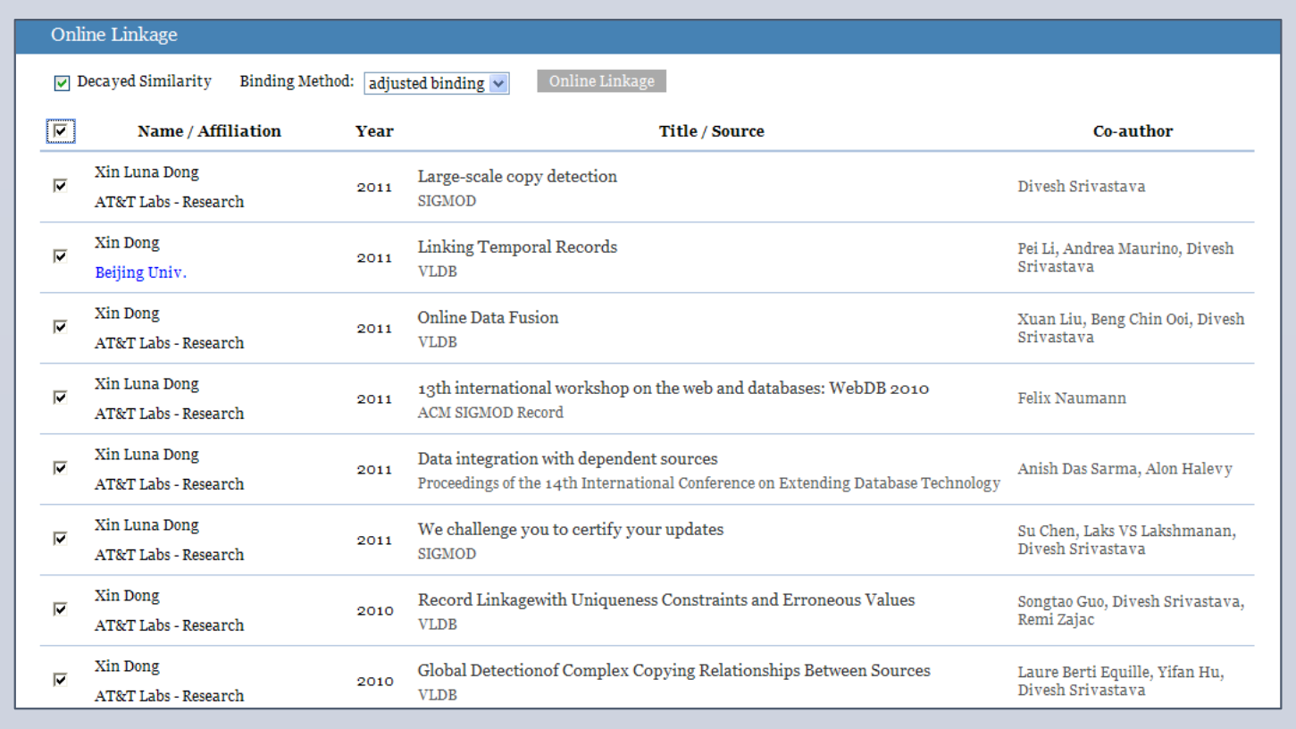
Xin Dong was in AT&T Labs - Research since 2008; the probability of moving from AT&T Labs - Research to University of Nebraska Lincoln in 2010 is 0.36

Xin Dong was in University of Nebraska Lincoln in 2010; the probability of moving from University of Nebraska Lincoln to AT&T Labs - Research in 2010 is 0.14

Chronos
1.0
0.8
0.6
0.4
0.2
0.0
0
5
10
15
20
25
30
Probability of Xin Luna Dong changing affiliation in 2 years.
Probability of Xin Luna Dong changing affiliation in 0 years.
2005
2011

Online linkage:

CHRONOS answers "what-if" questions by allowing the user to (1) select a subset of records, (2) change records' values (3) choose different linkage techniques and then compare the results.



Online Linkage

☑ Decayed Similarity Binding Method: adjusted binding Online Linkage

☑	Name / Affiliation	Year	Title / Source	Co-author
☑	Xin Luna Dong AT&T Labs - Research	2011	Large-scale copy detection SIGMOD	Divesh Srivastava
☑	Xin Dong Beijing Univ.	2011	Linking Temporal Records VLDB	Pei Li, Andrea Maurino, Divesh Srivastava
☑	Xin Dong AT&T Labs - Research	2011	Online Data Fusion VLDB	Xuan Liu, Beng Chin Ooi, Divesh Srivastava
☑	Xin Luna Dong AT&T Labs - Research	2011	13th international workshop on the web and databases: WebDB 2010 ACM SIGMOD Record	Felix Naumann
☑	Xin Luna Dong AT&T Labs - Research	2011	Data integration with dependent sources Proceedings of the 14th International Conference on Extending Database Technology	Anish Das Sarma, Alon Halevy
☑	Xin Luna Dong AT&T Labs - Research	2011	We challenge you to certify your updates SIGMOD	Su Chen, Laks VS Lakshmanan, Divesh Srivastava
☑	Xin Dong AT&T Labs - Research	2010	Record Linkagewith Uniqueness Constraints and Erroneous Values VLDB	Songtao Guo, Divesh Srivastava, Remi Zajac
☑	Xin Dong AT&T Labs - Research	2010	Global Detectionof Complex Copying Relationships Between Sources VLDB	Laure Berti Equille, Yifan Hu, Divesh Srivastava

Framework

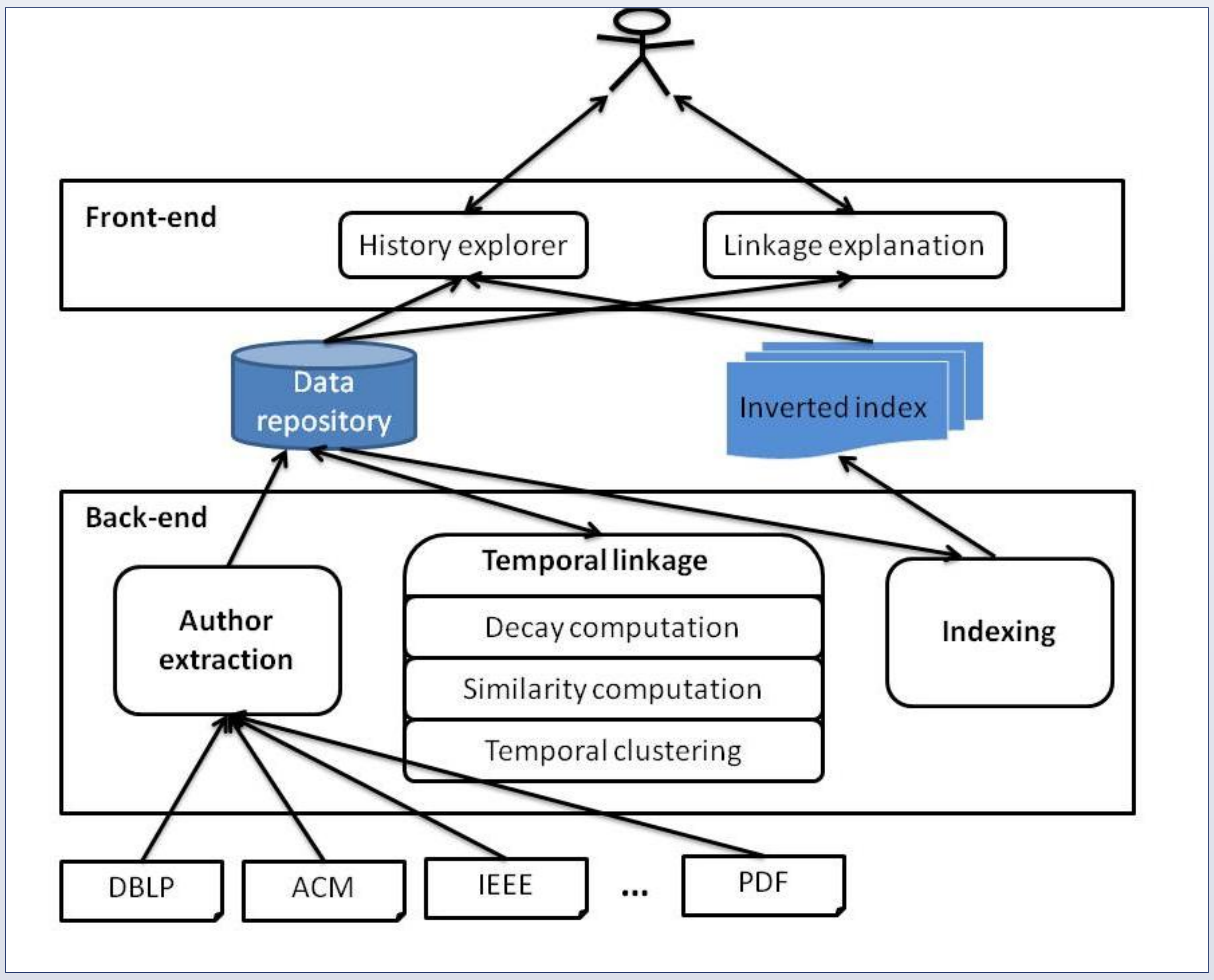
Data set:

Data extracted from more than 500 M publication entries in DBLP, ACM, Springer, IEEE etc..

Front-end

History explorer is the interface through which the user interacts with the system. It offers (1) author search by name, time period, and affiliation, (2) history tracing for each author, and (3) statistics view of the data.

Linkage explanation explains linkage decisions. It (1) shows the comparison of results from CHRONOS, from DBLP, and from BASIC, (2) explains the decision of a particular paper included in or excluded from the list of papers for a particular author, and (3) performs online temporal linkage and answers "what-if" questions.



Back-end

Author extraction: This component takes the DBLP data as input. For each paper, it extracts records about authors, including author name, paper title, conference, co-author, publication year from DBLP, and affiliation, email of the author from external sources (e.g., ACM, IEEE, journal websites, and PDF paper files).

Temporal linkage identifies author records that refer to the same real-world person. It contains three sub-components: *Decay computation*, *Similarity computation*, and *Temporal clustering*.

- **Decay computation**: One key idea of our temporal linkage algorithm is to apply time decay, which aims to capture the effect of time elapse on entity value evolution.
- **Similarity computation**: We compare a record with a cluster of records considering two aspects: (1) value consistency, and (2) continuity of the record with the cluster in time.
- **Temporal clustering:** We consider author records in time order for clustering and accumulate evidence overtime to enable global decision making.

Indexing builds an Inverted index for each identified real-world author. Each author is indexed by her names, affiliations, and also the years of her publications.

Reference

[1] P. Li, X. L. Dong, A. Maurino, and D. Srivastava. Linking Temporal Records. PVLDB, 4(11):956–967, 2011.

[2] O. Hassanzadeh, F. Chiang, H. C. Lee, and R. J. Miller. Framework for evaluating clustering algorithms in duplicate detection. PVLDB, 2(1):1282–1293, 2009.

Contact: pei.li@disco.unimib.it

Demo URL: http://siti-rack.siti.disco.unimib.it:8080/Chronos/