OLAP and Data Mining of Text Cubes for Aviation Safety Report Data Analysis

Jiawei Han, Cindy X. Lin, Lu Liu

University of Illinois Urbana/Champaign

in collaborations with Nikunj Oza and Ashok Srivastava

Support under NASA Project:

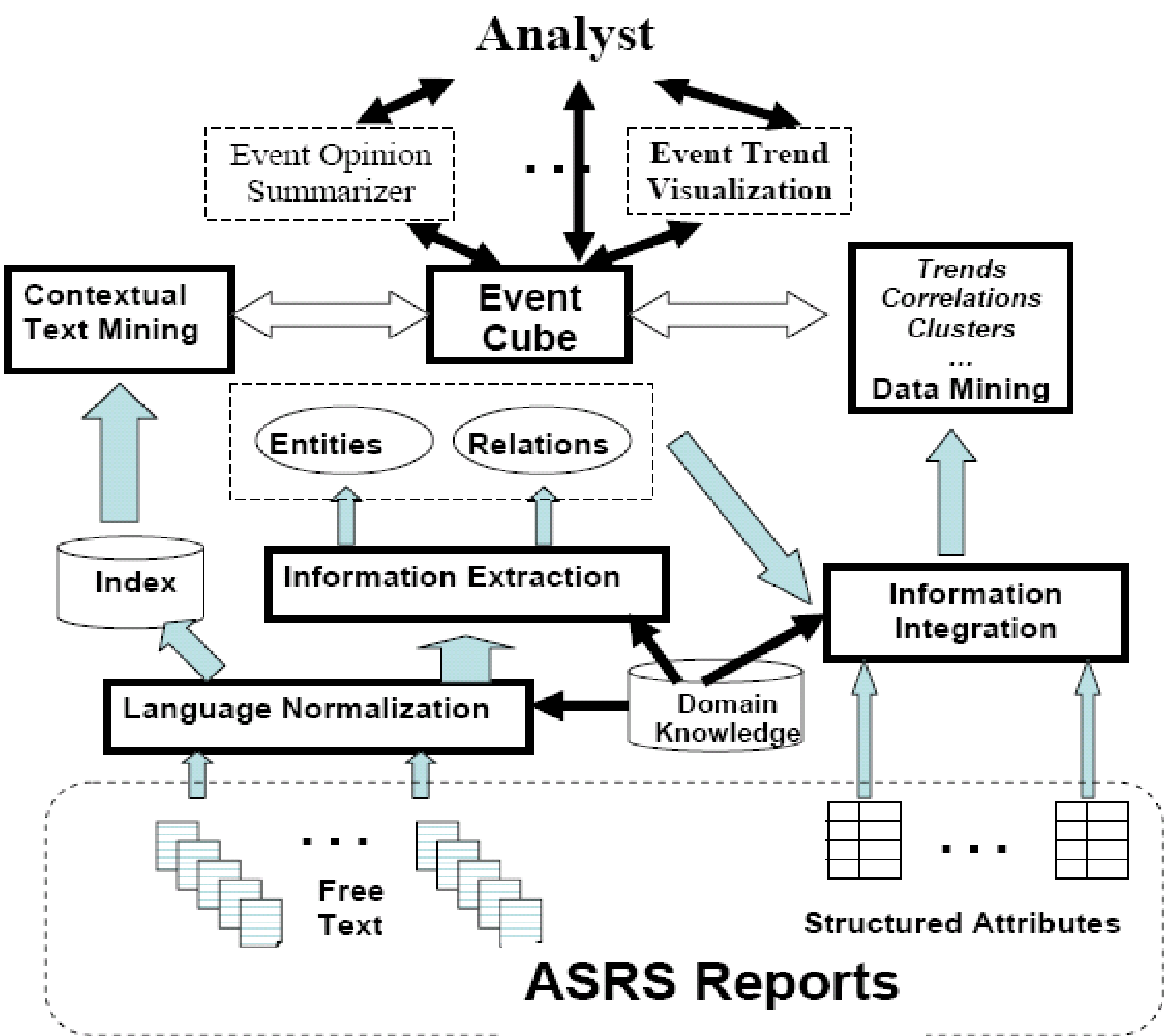
“Event Cube: An Organized Approach for Mining and Understanding Anomalous Aviation Events”

August 10, 2009

Outline

- **The Event Cube Project**
- **Multi-Dimensional Analysis of Text Data**
  - **Text Cube: Basic IR measure for MD-Text Data**
  - **Topic Cube: Topic modeling of MD-Text Data**
  - **Comparing Cube: Topic modeling with user-based discriminative analysis**
- **iNextCube: Towards Information Network Analysis in Event Cubes**

Event Cube for Multi-Dimensional Text Mining in ASRS Datasets



Research Team on Event Cubes

- **University of Illinois at Urbana-Champaign**
  - **Jiawei Han:** Data mining, data cube and database systems
  - **Chengxiang Zhai:** Text mining, information retrieval
  - **Students: Bolin Ding** (Sequence mining), **Samson Hauguel** (Automatics concept hierarchy building), **Cindy X. Lin** (Event cube architect, text cube/search), **Lu Liu** (Text mining), **Duo Zhang** (Text mining, topic cube), **Bo Zhao** (Multidimensional text analysis), **Feida Zhu** (Motif mining)
- **University of Texas at Dallas**
  - **Latifur Khan:** Data mining, text mining
  - **Vincent Ng:** Natural language analysis, text mining
  - **Bhavani Thuraisingham:** Information security, data mining, text mining
  - **Students: Md. Arshad Ul Abedin** (Cause analysis), **Salim Ahmed** (Anomaly detection), **Greg Hellings** (Language normalization), **Qing Chen** (Language normalization)
- **Boeing: Phantom Works (Collaborator)**
  - **Anne Kao:** Data Mining, aviation safety analysis

Outline

- The Event Cube Project
- Multi-Dimensional Analysis of Text Data
  - Text Cube: Basic IR measure for MD-Text Data
  - Topic Cube: Topic modeling of MD-Text Data
  - Comparing Cube: Topic modeling with user-based discriminative analysis
- iNextCube: Towards Information Network Analysis in Event Cubes

Analysis of Multi-Dimensional Text Data

- ASRS Dataset: A typical multi-dimensional text database
- Analysis of (multi-dimensional) relational database:
  - Data cube and OLAP (online analytical processing): driving engine in database industry
- Analysis of multi-dimensional text data
  - Integration of data cube and information retrieval (IR)
  - Text Cube, Topic Cube, Comparing Cube
- Text cube
  - Cindy X. Lin, Bolin Ding, Jiawei Han, Nikunj C.Oza, Ashok N,Srivastava, Bo Zhao and Feida Zhu, "*Text Cube: Computing IR Measures for Multidimensional Text Database Analysis*", journal version (with NASA researchers) submitted to IEEE Trans. on Knowledge and Data Engineering, original version in Proc. 2008 Int. Conf. on Data Mining (ICDM'08), Pisa, Italy, Dec. 2008,
- Topic cube:
  - Duo Zhang, Chengxiang Zhai and Jiawei Han, "*Topic Cube: Topic Modeling for OLAP on Multidimensional Text Databases*", Proc. 2009 SIAM Int. Conf. on Data Mining (SDM'09), Sparks, Nevada, Apr. 2009 (Best of SDM'09), Journal version (with NASA researchers) invited to the special issue of SDM'09

Problem: ASRS Dataset

ACN	Time	Weather	…	Anomaly Event	Flight Safety Report
10001	2006	Ice	… …	Excursion: Runway	…the aircraft began to slide left and right…
10002	2007	Rain	… …	Excursion: Runway	… the adjacent pavement filled with water…
10003	2008	Rain	… …	Inflight Encounter: Birds	…a flock of seagulls on the circle line of the runway...
……	……	……	……	……	……

Traditional Methods:

1. Data Cube is a powerful tool to support OLAP for structured multidimensional categorical data (e.g., the part in red frame)

2. IR (Information Retrieval) techniques help analyze unstructured flat free text data (e.g. the part in green frame)

Motivation:

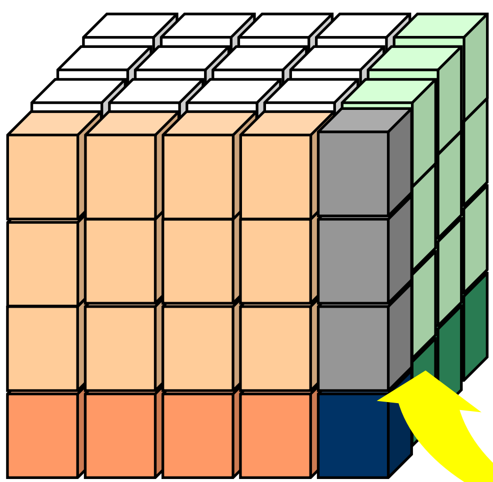
How to deal with heterogeneous dataset like the above ASRS (Aviation Safety Reporting System) dataset, which has both structured and unstructured information ?

General Idea

- Heterogeneous: categorical attributes + unstructured text

ACN	Time	Location	Place	Environment	… …	Event	Report
							Text data

- How to combine?
- Our solution:



Cube: Categorical Attributes

Measure

Term/Topic	Weight
T1	W1
T2	W2
T3	W3
…	…

Text/Topic Model: Unstructured Text

Text Cube: General Idea

Dimension

Use Structured Categorical Information

Dimension Hierarchy: As traditional OLAP cube, each dimension consists of multi attributes and is organized as a tree or DAG.
Four operations: roll-up, drill-down, slice, dice.

+

Measure

Summary Statistics on Unstructured Text

Text Cube

A novel data cube, where two kinds of information can mutually enhance the knowledge discovery of each other.

Preprocessing on text:
Step 1: Utilize WordNet to stem terms
Step 2: use TF-IDF to weight terms, keep the top k terms with highest weights as Topic Term
Step 3: Count TF and IDF of Topic Terms.

Measure Supported
1. TF: term frequency 2. IV: inverted index

Term Hierarchy
1. semantic levels of terms and their relationships
2. given by domain experts.

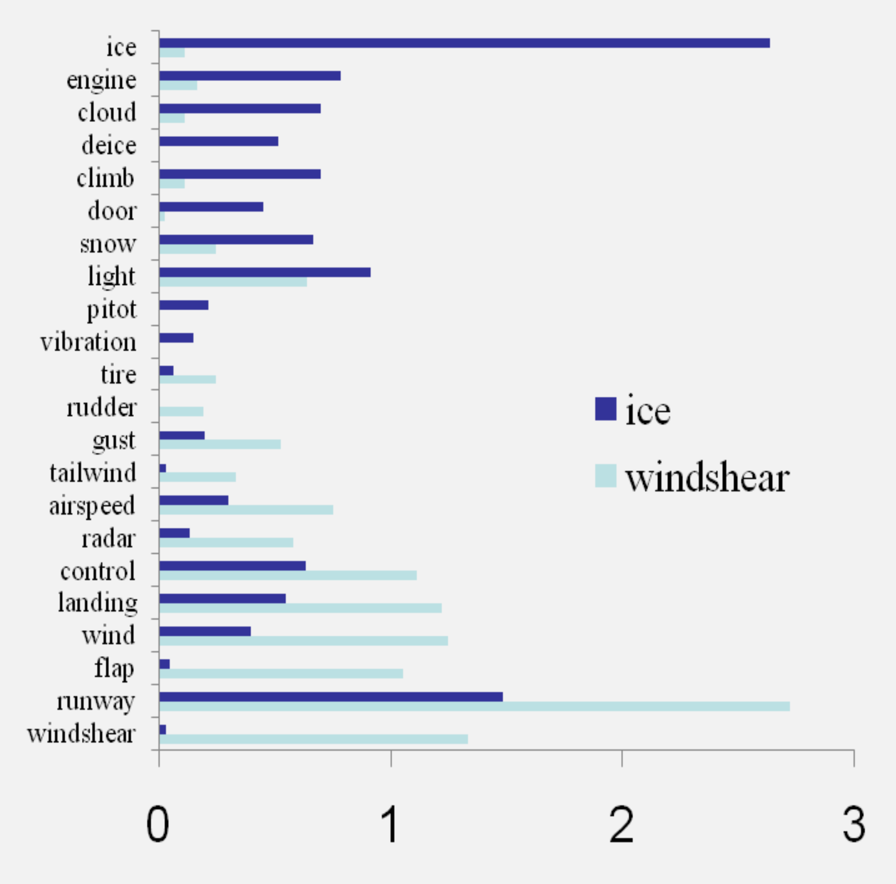
Term Level
An arbitrary cut on term hierarchy tree

Two Operation
1. push-down : replace one tree node by its children node
2. pull-up : the reverse of push-down

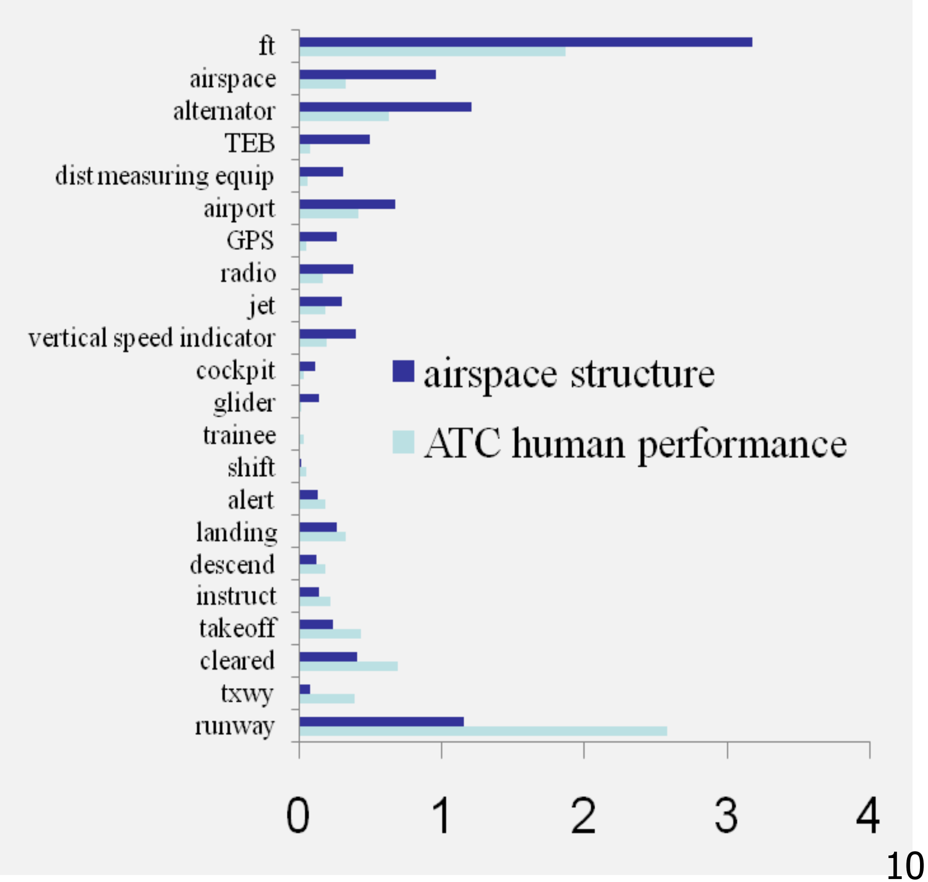
Text Cube: Some Experimental Results

Interesting Result: (avgTF = TF / count)

Compare avgTF under different "*Environment: Weather Elements*"



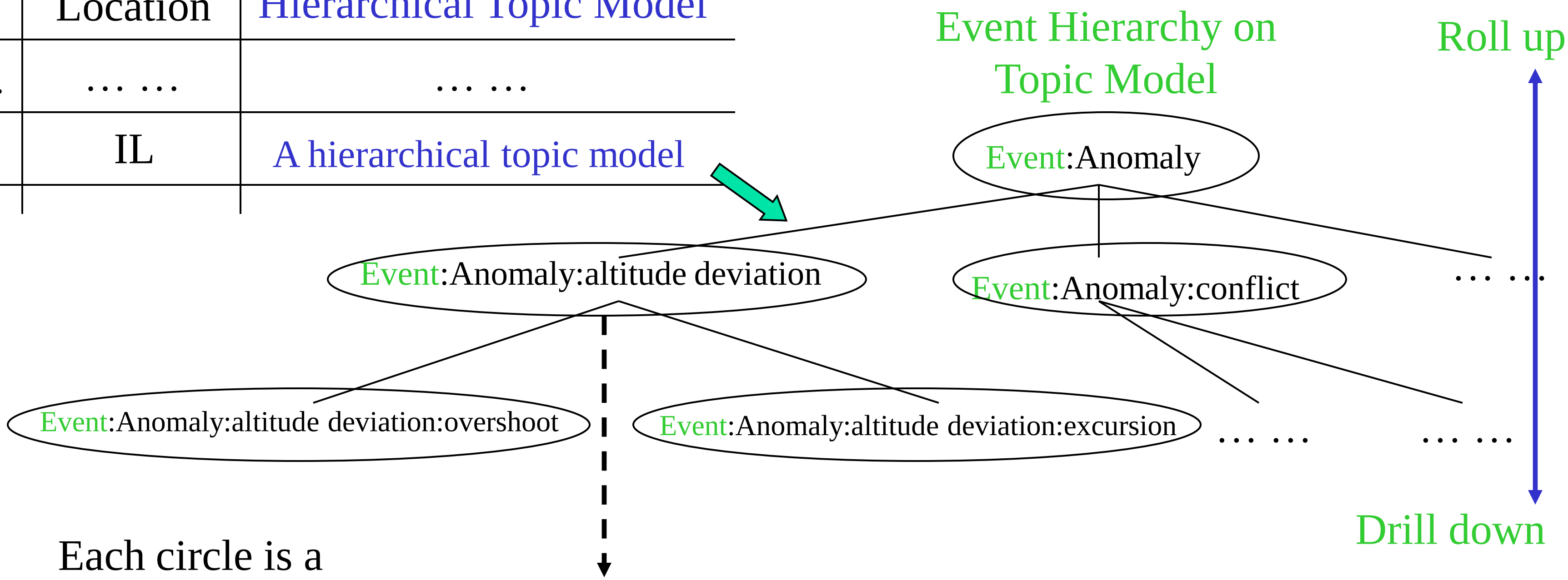
Compare avgTF under different "*Supplementary: Problem Areas*"



Topic Cube: Multidimensional Text Analysis by Topic Modeling

Unstructured Text ➔ Topic Model

Time	Location	Hierarchical Topic Model
… …	… …	… …
2007	IL	A hierarchical topic model

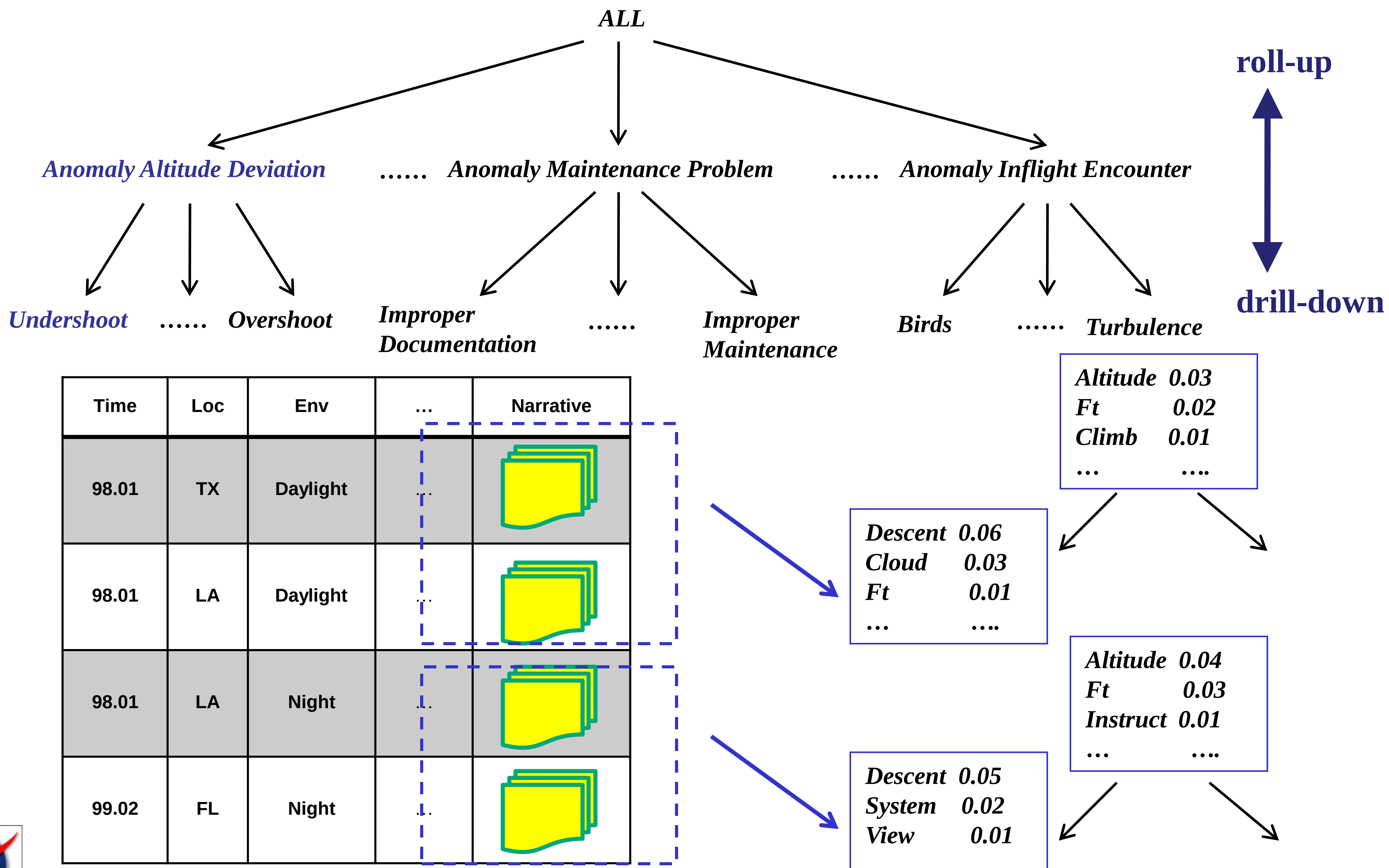


Each circle is a Topic Model For example:

Term	Probability
altimeter	0.01
leveling	0.008
leveloff	0.007
supplemental	0.006
…	…

Clustering,
Classification,
…
using topic models

Topic Cube: Cubing Algorithm



Time	Loc	Env	…	Narrative
98.01	TX	Daylight	…	
98.01	LA	Daylight	…	
98.01	LA	Night	…	
99.02	FL	Night	…	

Topic Cube: Experimental Results

Topic Content Comparison

---- landing without clearance

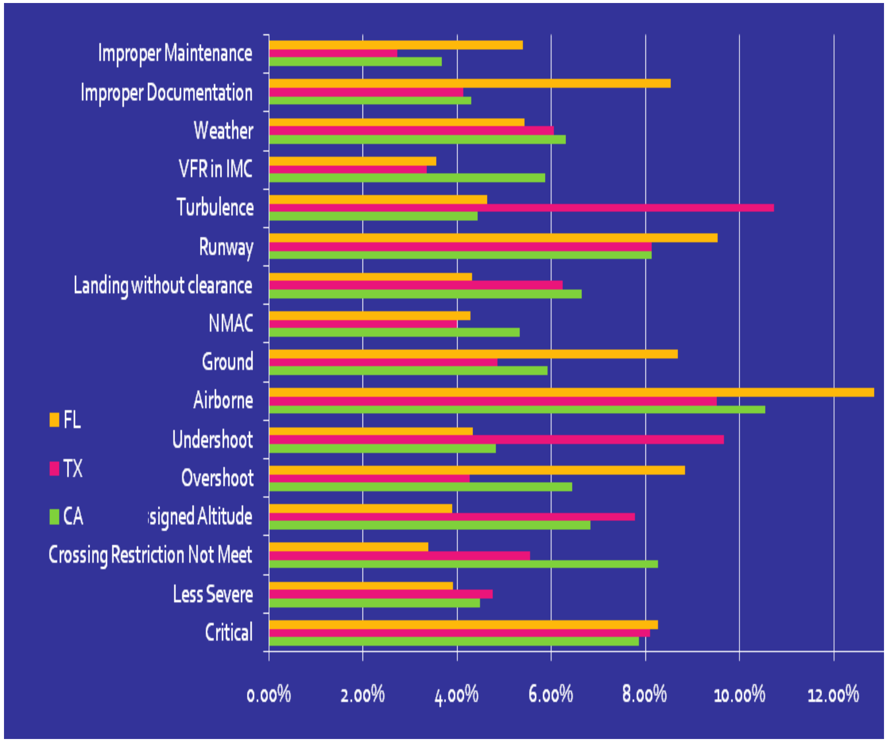
Context	Word	p(w|θ)
daylight	Tower	0.075
	Pattern	0.061
	Final	0.060
	Runway	0.053
	Land	0.052
	Downwind	0.039
night	Tower	0.035
	Runway	0.029
	Light	0.027
	Instrument Landing System	0.015
	Beacon	0.014

…WINDS ALOFT AT PATTERN ALT OF 1000 FT MSL, WERE MUCH STRONGER AND A DIRECT XWIND. NEEDLESS TO SAY, THE PATTERNS AND LNDGS WERE DIFFICULT FOR MY STUDENT AND THERE WAS LIGHT TURB ON THE DOWNWIND…

…I LISTENED TO HWD ATIS AND FOUND THE TWR CLOSED AND AN ANNOUNCEMENT THAT THE HIGH INTENSITY LIGHTS FOR RWY 28L WERE INOP. BROADCASTING IN THE BLIND AND LOOKING FOR THE TWR BEACON AND LOW INTENSITY LIGHTS AGAINST A VERY BRIGHT BACKGROUND CLUTTER OF STREET LIGHTS, ETC…

Topic Coverage Comparison

We can compare topics in different context by comparing the coverage of these topics in different cells (e.g., location comparison)



Text Summary — Expect More Discriminative Results

Input Format

1. Specify a cell of Text Cube, e.g., [Year] = '2008', other dimensions = 'All', which results in the table in page 'Problem : ASRS dataset'.

2. Specify a comparing dimension of the cell and several comparing values, e.g. we want to compare the text summaries when [Anomaly Event] = 'inflight encounter : birds'(topic 1), 'ground encounters : animal' (topic 2) and 'excursion : runway' (topic 2);

Output Format

Showing the top 10 words according to their probabilities in the resulting word distribution

topic1	topic2	topic3
Engine	Aircraft	Runway
Bird	Runway	Aircraft
Aircraft	Engine	Land
Normal	Takeoff	Time
Damage	Tire	Takeoff
Runway	Brake	ft
Strike	Damage	Approach

The text summaries of the three topics are too similar. We expect more discriminative results.

Improvement : Comparing Cube

General Idea

1. The user specifying a comparing dimension implies that he/she expects more information about the comparing dimension in the results. But Topic Cube and other traditional topic extracting methods ignored such kind of 'user intention'.

2. Text Cube may introduce irrelevant background information, e.g., 'Runway' and 'Aircraft' , but we do not expect these 'too common' words.

Improvement

Comparing Cube utilizes a modified PLSA algorithm to calculate the related extent between one sentence and the comparing dimension, and only remains relevant sentences in the flight reports.

Modified PLSA

$$Z_{i,j}(k) = \frac{\pi_i(k) \sum_w c(j,w)\theta_k(w)}{\sum_{k'} \pi_i(k') \sum_w c(j,w)\theta_k(w)}$$

$$\pi_i(k) = \frac{\sum_j Z_{i,j}(k)}{\sum_{k'} \sum_j Z_{i,j}(k')}$$

$$\theta_k(j) = \frac{\sum_i Z_{i,j}(k)}{\sum_{j'} \sum_i Z_{i,j'}(k)}$$

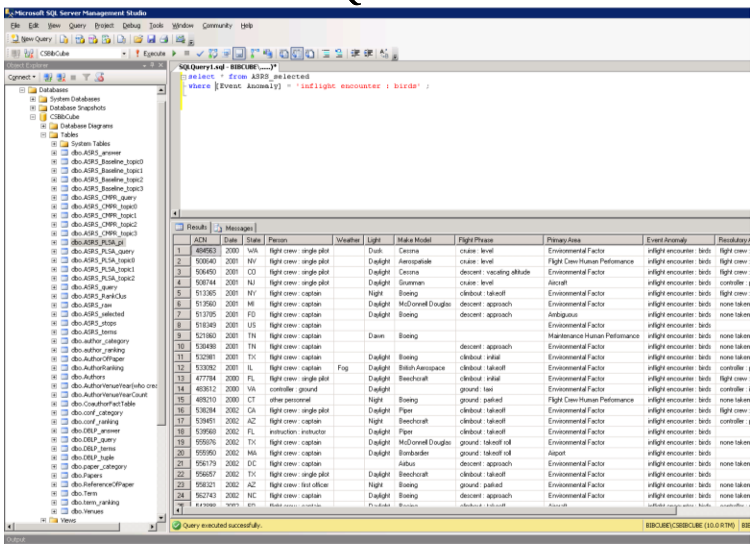
Improved Results

topic1	topic2	topic3
Bird	Animal	Runway
Runway	Gate	FT
Dead	Brake	Land
Fly	Deer	Notam
Seagull	Mechanical	Slide
Aircraft	Cleared	Brake
Ocean	Bird	Turn

System Implementation

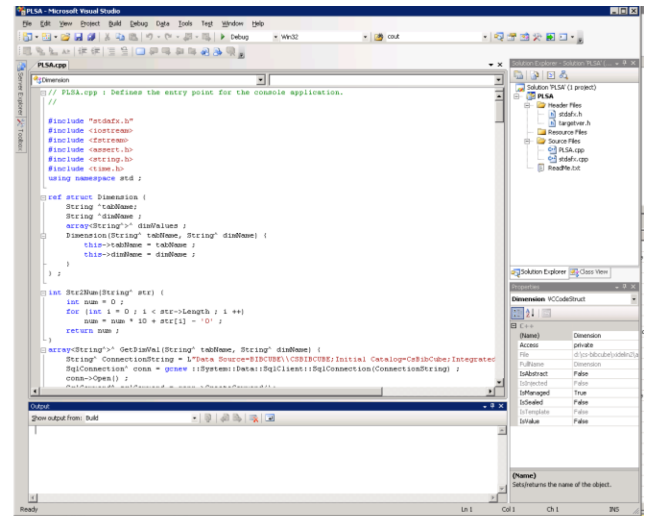
Step 1: Data Storage

---- Microsoft SQL Server Database



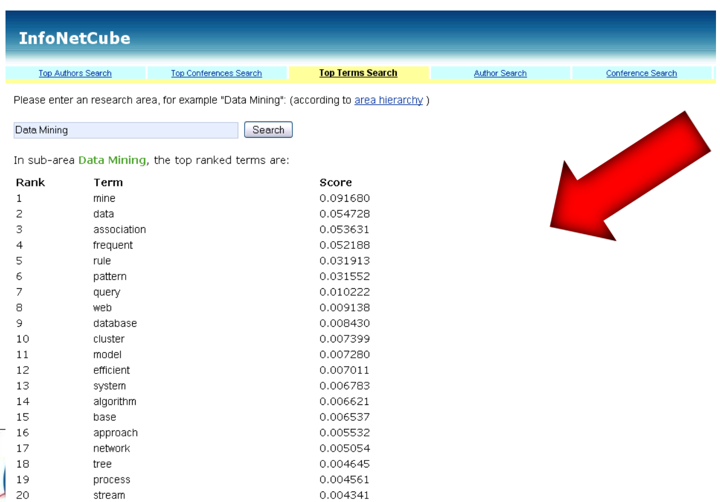
Step 2: Computation

---- Microsoft Visual Studio



Step 3: Communication

----Common Language Runtime Support



Rank	Term	Score
1	mine	0.091680
2	data	0.054728
3	association	0.053631
4	frequent	0.052188
5	rule	0.031913
6	pattern	0.031552
7	query	0.010222
8	web	0.009138
9	database	0.008430
10	cluster	0.007399
11	model	0.007280
12	efficient	0.007011
13	system	0.006783
14	algorithm	0.006621
15	base	0.006537
16	approach	0.005532
17	network	0.005054
18	tree	0.004645
19	process	0.004561
20	stream	0.004341

Step 4: Web User Interface

---- Microsoft C #

Step 5: Visualization

---- Flash, Flare

Text/Topic/Comparing Cube: A Powerful Framework to Be Fully Developed

- Document classification and clustering in each cell
  - E.g. cluster the documents of one cell into anomaly events
- Analyze the importance of anomaly events in each cell
  - E.g. in Place A and Time B, which kind of anomaly event occurs the most?
- Mining correlation between anomaly events and contexts
  - E.g. in what kind of weather condition, the flight has the problem of <u>landing without clearance</u>?
- Cause analysis based on the text cube
  - E.g. what are the main factors of each anomaly event in different situations?
- …

Mining Repetitive Gapped Subsequence in Text

- Mining ASRS dataset
  - Anomaly1 = aircraft equipment problem : critical
  - Anomaly2 = inflight encounter : weather
  - Anomaly3 = conflict : nmac

Pattern	Support		
	Anomaly 1	Anomaly 2	Anomaly 3
LNDG UNEVENTFUL	11	0	0
LANDED WITHOUT INCIDENT	12	0	0
SHUT DOWN ENG	12	0	0
VISIBILITY FOG	0	13	0
CEILING VISIBILITY	0	15	0
DOWNWIND RWY	0	0	12
SAW OTHER ACFT	0	0	10
CLRED FOR RWY	0	0	44
TOOK EVASIVE ACTION	0	0	44
SUPPLEMENTAL FROM	17	10	31
CALLBACK WITH REVEALED FOLLOWING	37	13	24
CALLBACK WITH REVEALED FOLLOWING HAT	13	0	0

Utility of patterns

Correlation between patterns (word sequences) and anomalies

Describe/explain anomalies with patterns

A generic approach

Can be also applied in other kinds of "sequences", like *cockpit switch sequences*

Extension

Using them as features for classification or clustering

"Efficient Mining of Closed Repetitive Gapped Subsequences from a Sequence Database", Proc. 2009 Int. Conf. on Data Engineering (ICDE'09), with Ashok and Nikunj, submitted to IEEE Trans. on Knowledge and Data Engineering, 2009

Outline

- **The Event Cube Project**
- **Multi-Dimensional Analysis of Text Data**
  - **Text Cube: Basic IR measure for MD-Text Data**
  - **Topic Cube: Topic modeling of MD-Text Data**
  - **Comparing Cube: Topic modeling with user-based discriminative analysis**
- **iNextCube: Towards Information Network Analysis in Event Cubes**

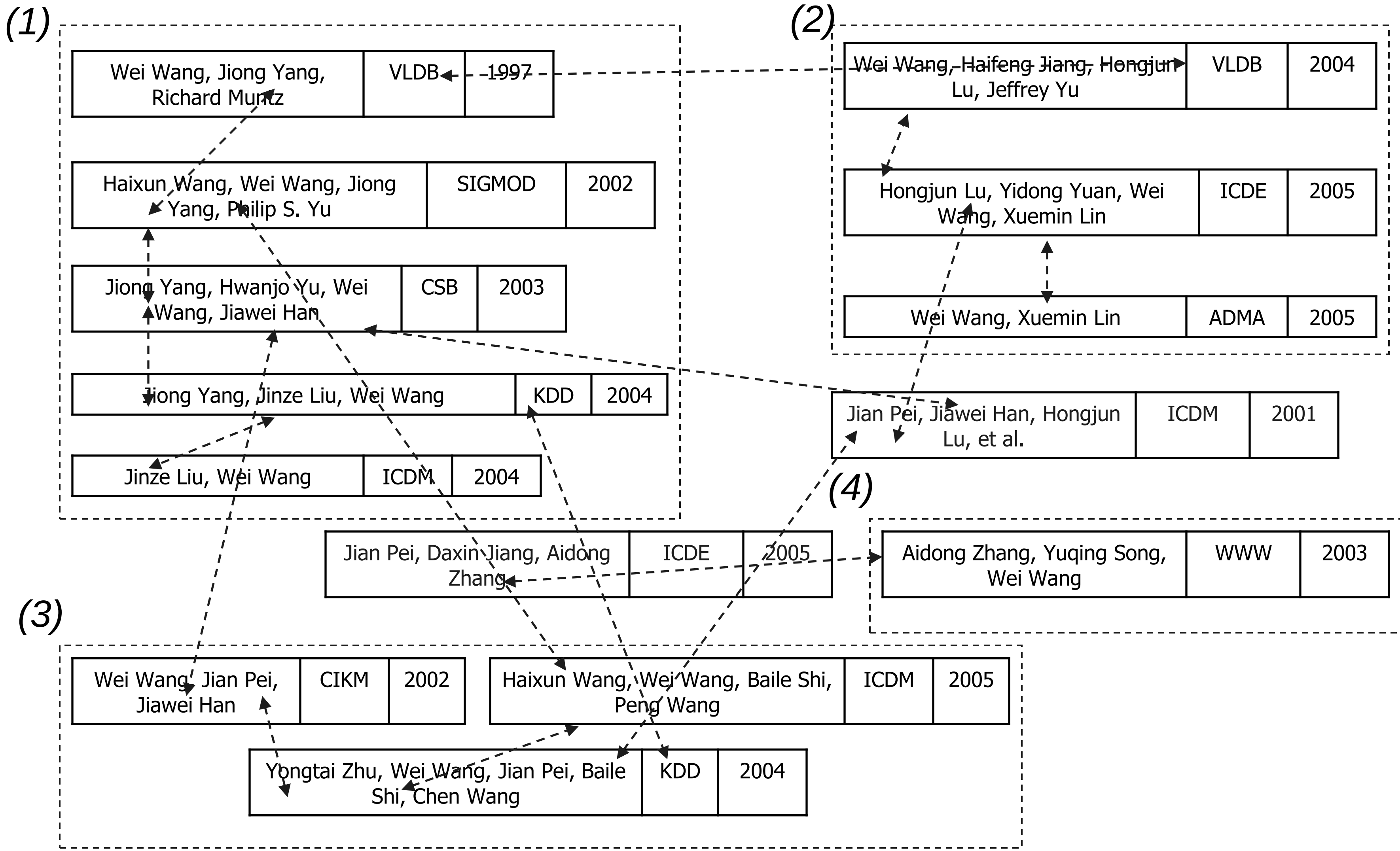
Towards Information Network Analysis in ASRS Report Analysis

- Knowledge is power, but knowledge is hidden in massive information networks
- Information network analysis is powerful at uncovering knowledge hidden in massive links and networks
  - Distinguish identical names (info. Integration)
  - Validation of conflict facts (veracity analysis)
  - Rank-based clustering for hierarchy discovery
- We are constructing iNextCube for integration of multidimensional text analysis and information network analysis
- iNextCube will be demo. In VLDB’09, Aug., Lyon, France

Object Reconciliation by Link Analysis

- Link makes entity cross-checking and validation easy
- Object reconciliation vs. object distinction
- Object distinction: People/objects do share names
  - In AllMusic.com, 72 songs and 3 albums named "Forgotten" or "The Forgotten"
  - In DBLP, 141 papers are written by at least 14 "Wei Wang"
- Distinct: Object distinction by information network analysis
  - X. Yin, J. Han, and P. S. Yu, "Object Distinction: Distinguishing Objects with Identical Names by Link Analysis", ICDE'07

Entity Distinction: The "Wei Wang" Challenge in DBLP



(1) Wei Wang at UNC *(2) Wei Wang at UNSW, Australia*

(3) Wei Wang at Fudan Univ., China *(4) Wei Wang at SUNY Buffalo*

DISTINCT: Analysis Methodology

- Measure similarity between references
  - Link-based similarity: Linkages between references
    - References to the same object are more likely to be connected
  - Neighborhood similarity
    - Neighbor tuples of each reference can indicate similarity between their contexts
- Self-boosting: Training using the “same” bulky data set
- Reference-based clustering
  - Group references according to their similarities

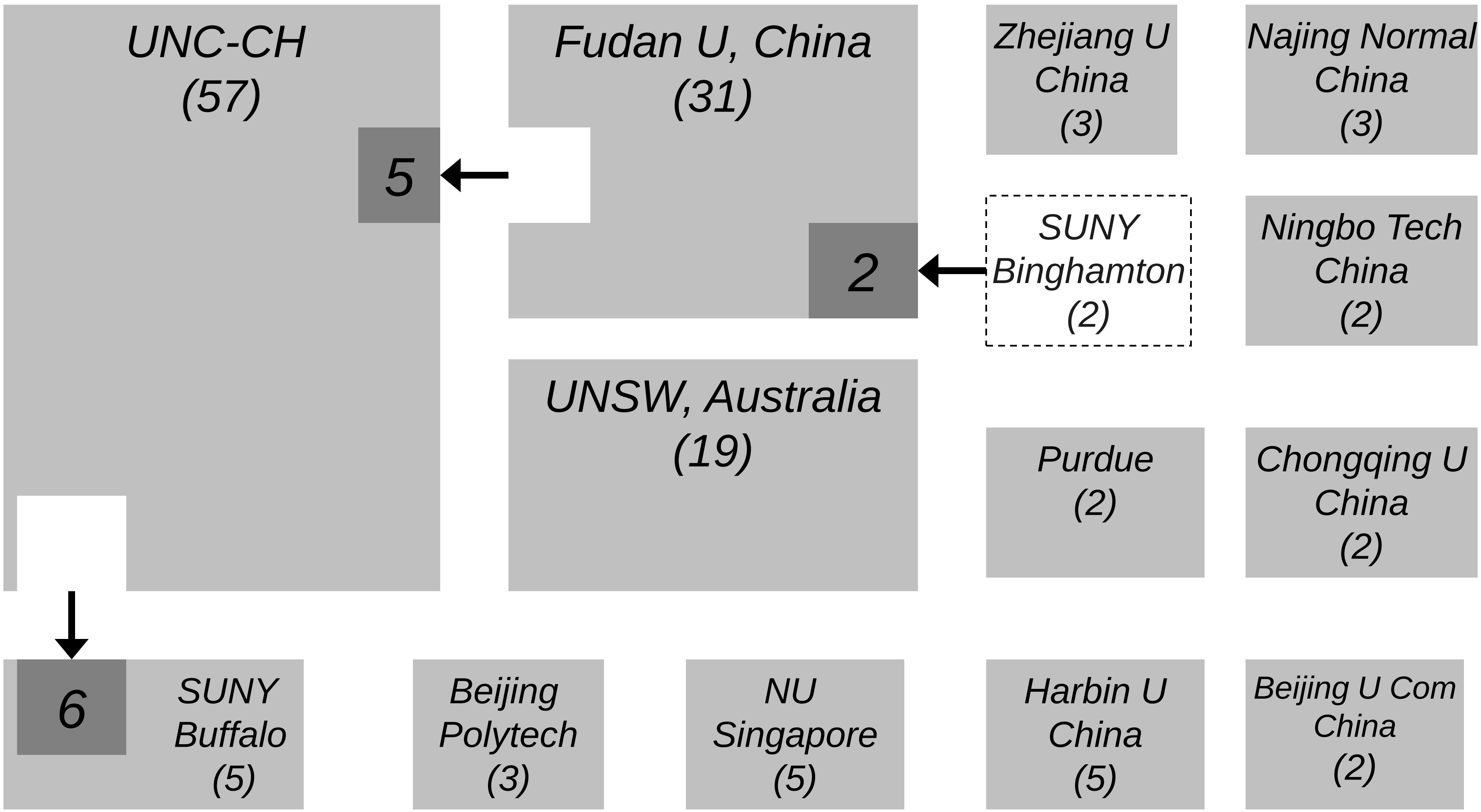
Training with the “Same” Data Set

- Build a training set automatically
  - Select distinct names, e.g., Johannes Gehrke
  - The collaboration behavior within the same community share some similarity
  - Training parameters using a typical and large set of “unambiguous” examples
- Use SVM to learn a model for combining different join paths
  - Each join path is used as two attributes (with link-based similarity and neighborhood similarity)
  - The model is a weighted sum of all attributes

Experiments with DBLP Data

Name	*#author*	*#ref*	*accuracy*	*precision*	*recall*	*f-measure*
Hui Fang	3	9	1.0	1.0	1.0	1.0
Ajay Gupta	4	16	1.0	1.0	1.0	1.0
Joseph Hellerstein	2	151	0.81	1.0	0.81	0.895
Rakesh Kumar	2	36	1.0	1.0	1.0	1.0
Michael Wagner	5	29	0.395	1.0	0.395	0.566
Bing Liu	6	89	0.825	1.0	0.825	0.904
Jim Smith	3	19	0.829	0.888	0.926	0.906
Lei Wang	13	55	0.863	0.92	0.932	0.926
Wei Wang	14	141	0.716	0.855	0.814	0.834
Bin Yu	5	44	0.658	1.0	0.658	0.794
average			0.81	0.966	0.836	0.883

Distinguishing Different "Wei Wang"s



Truth Validation by Information Network Analysis

- Xiaoxin Yin, Jiawei Han, Philip S. Yu, "Truth Discovery with Multiple Conflicting Information Providers on the Web", KDD'07
- The trustworthiness problem of the web (according to a survey):
  - 54% of Internet users trust news web sites most of time
  - 26% for web sites that sell products
  - 12% for blogs
- TruthFinder: Truth discovery on the Web by link analysis
  - Among multiple conflict results, can we automatically identify which one is likely the true fact?
- Veracity (conformity to truth):
  - Given a large amount of conflicting information about many objects, provided by multiple web sites (or other information providers), how to discover the true fact about each object?

Conflicting Information on the Web

- Different websites often provide conflicting info. on a subject, e.g., Authors of *“Rapid Contextual Design”*

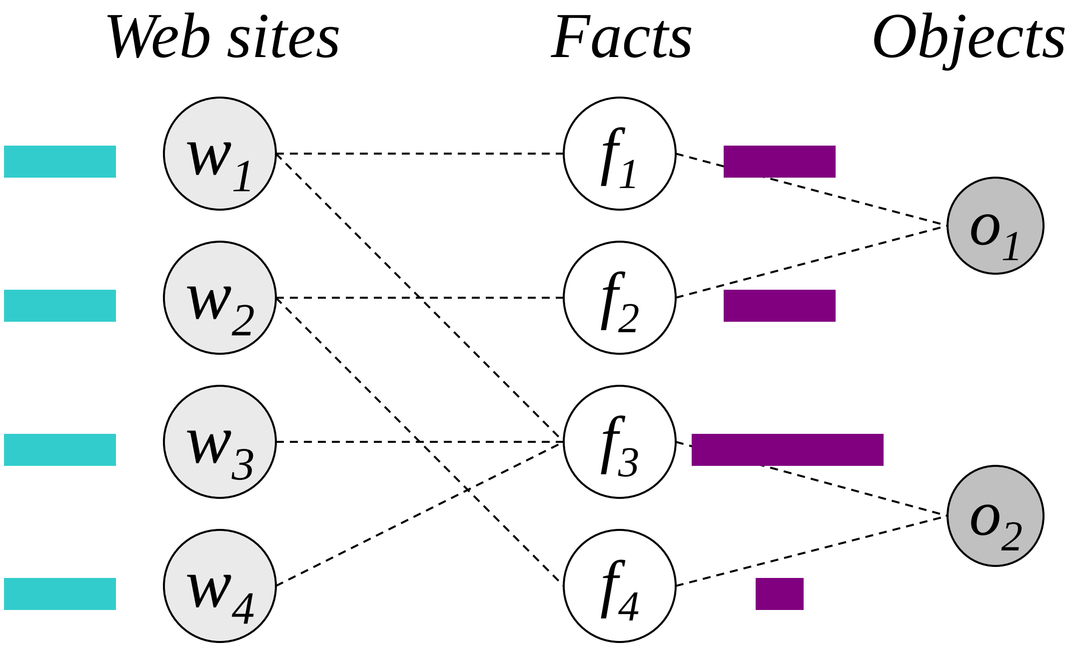
Online Store	*Authors*
Powell’s books	Holtzblatt, Karen
Barnes & Noble	Karen Holtzblatt, Jessamyn Wendell, Shelley Wood
A1 Books	Karen Holtzblatt, Jessamyn Burns Wendell, Shelley Wood
Cornwall books	Holtzblatt-Karen, Wendell-Jessamyn Burns, Wood
Mellon’s books	Wendell, Jessamyn
Lakeside books	WENDELL, JESSAMYNHOLTZBLATT, KARENWOOD, SHELLEY
Blackwell online	Wendell, Jessamyn, Holtzblatt, Karen, Wood, Shelley

Basic Heuristics for Problem Solving

1. There is usually only one true fact for a property of an object
2. True fact appears to be the same or similar on different web sites
   - E.g., "Jennifer Widom" vs. "J. Widom"
3. False facts on different web sites are less likely to be the same or similar
   - False facts are often introduced by random factors
4. A web site that provides mostly true facts for many objects will likely provide true facts for other objects

Inference on Trustworthiness

- Inference of web site trustworthiness & fact confidence



True facts and trustable web sites will become apparent after some iterations

Experiments: Finding Truth of Facts

- Determining authors of books
  - Dataset contains 1,265 books listed on abebooks.com
  - We analyze 100 random books (using book images)

Case	***Voting***	***TruthFinder***	***Barnes & Noble***
Correct	71	85	64
Miss author(s)	12	2	4
Incomplete names	18	5	6
Wrong first/middle names	1	1	3
Has redundant names	0	2	23
Add incorrect names	1	5	5
No information	0	0	2

Experiments: Trustable Info Providers

- Finding trustworthy information sources
  - Most trustworthy bookstores found by TruthFinder vs. Top ranked bookstores by Google (query “bookstore”)

TruthFinder

Bookstore	***trustworthiness***	***#book***	***Accuracy***
TheSaintBookstore	0.971	28	0.959
MildredsBooks	0.969	10	1.0
Alphacraze.com	0.968	13	0.947

Google

Bookstore	***Google rank***	***#book***	***Accuracy***
Barnes & Noble	1	97	0.865
Powell’s books	3	42	0.654

RankClus: Integration Ranking and Clustering

- Ranking and clustering each can provide general views over info-net
- Ranking globally without considering clusters → dumb
  - Ranking DB and Architecture Conferences. Together?
- Clustering authors in one huge cluster without distinction?
  - Dull to view thousands of authors
- RankClus: Integrate clustering with ranking
  - Conditional ranking relative to clusters
  - Uses highly ranked objects to improve clusters
- Quality of clustering and ranking are mutually enhanced
- Y. Sun, J. Han, et al., "*RankClus: Integrating Clustering with Ranking for Heterogeneous Information Network Analysis*", EDBT'09.

Global Ranking vs. Within-Cluster Ranking in a Toy Example

- Two areas: 10 conferences and 100 authors in each area

Table 1: A set of conferences from two research areas

DB/DM	{SIGMOD, VLDB, PODS, ICDE, ICDT, KDD, ICDM, CIKM, PAKDD, PKDD}
HW/CA	{ASPLOS, ISCA, DAC, MICRO, ICCAD, HPCA, ISLPED, CODES, DATE, VTS }

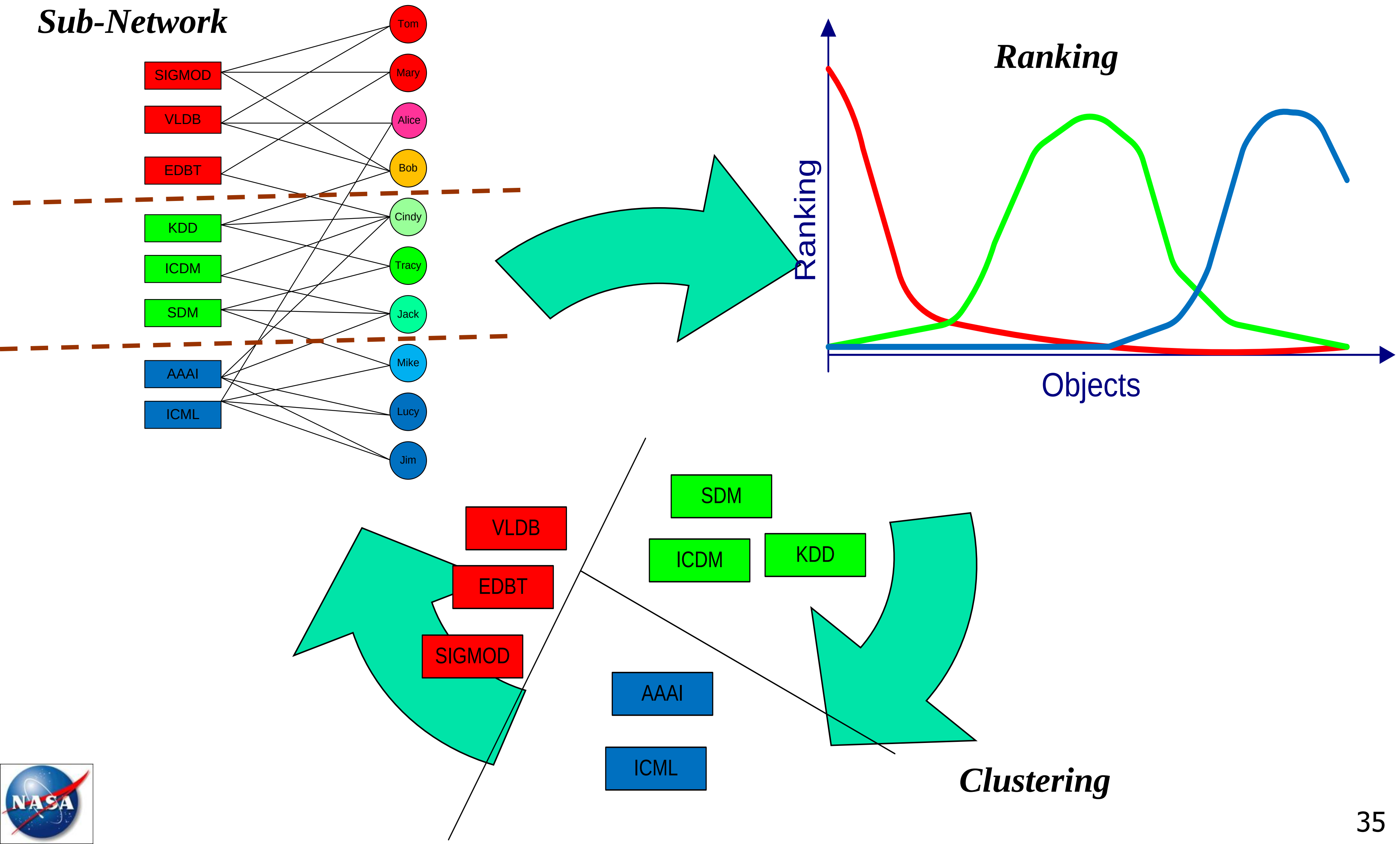
Table 2: Top-10 ranked conferences and authors in the mixed conference set

Rank	Conf.	Rank	Authors
1	DAC	1	Alberto L. Sangiovanni-Vincentelli
2	ICCAD	2	Robert K. Brayton
3	DATE	3	Massoud Pedram
4	ISLPED	4	Miodrag Potkonjak
5	VTS	5	Andrew B. Kahng
6	CODES	6	Kwang-Ting Cheng
7	ISCA	7	Lawrence T. Pileggi
8	VLDB	8	David Blaauw
9	SIGMOD	9	Jason Cong
10	ICDE	10	D. F. Wong

able 3: Top-10 ranked conferences and authors B/DM set

Rank	Conf.	Rank	Authors
1	VLDB	1	H. V. Jagadish
2	SIGMOD	2	Surajit Chaudhuri
3	ICDE	3	Divesh Srivastava
4	PODS	4	Michael Stonebraker
5	KDD	5	Hector Garcia-Molina
6	CIKM	6	Jeffrey F. Naughton
7	ICDM	7	David J. DeWitt
8	PAKDD	8	Jiawei Han
9	ICDT	9	Rakesh Agrawal
10	PKDD	10	Raghu Ramakrishnan

Algorithm Framework - Illustration

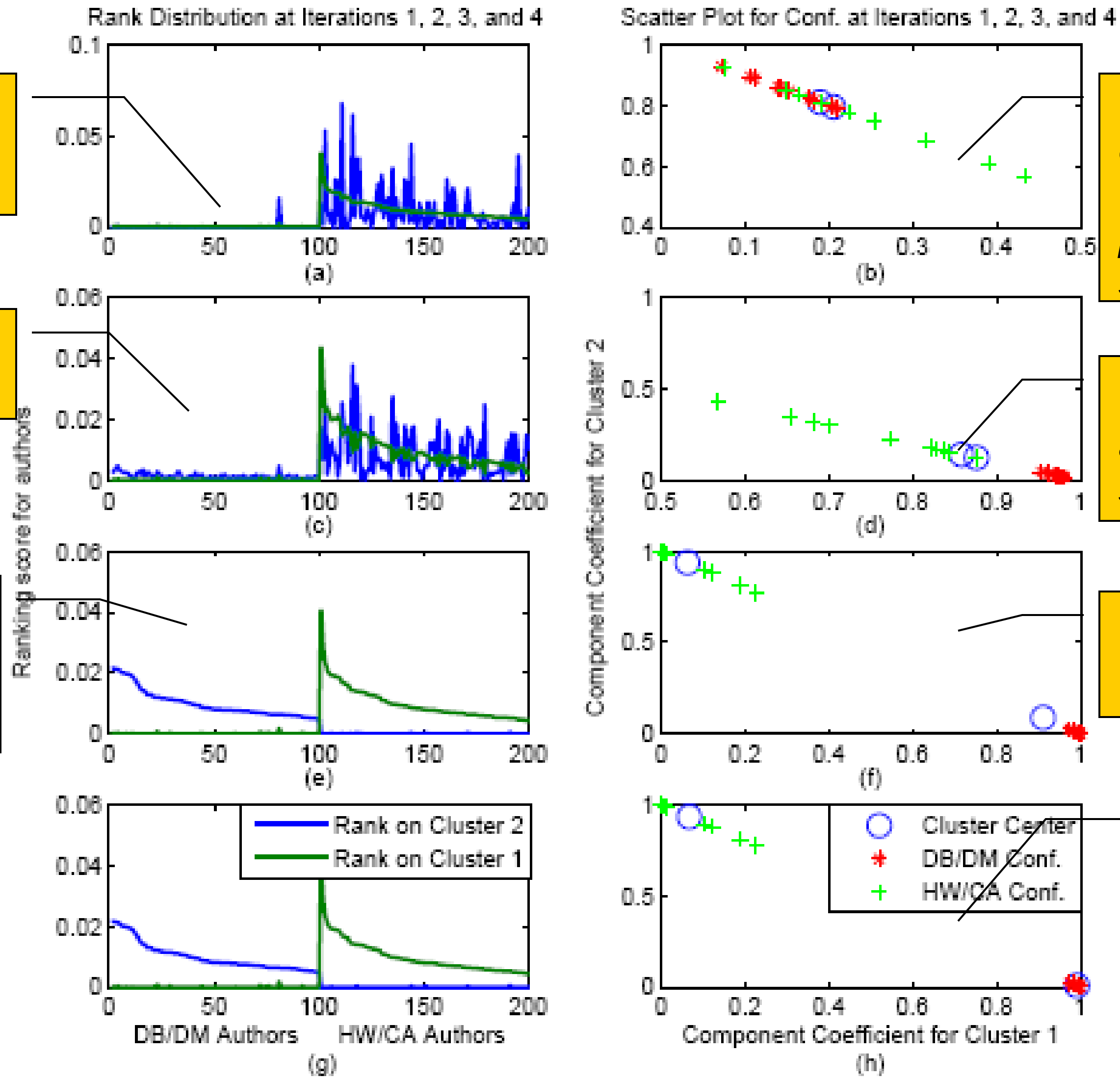


A Running Case Illustration for 2-Area Conf-Author Network

Initially, ranking distributions are mixed together

Improved a little

Improved significantly



Two clusters of objects mixed together, but preserve similarity somehow

Two clusters are almost well separated

Well separated

Stable

Case Study: Dataset: DBLP

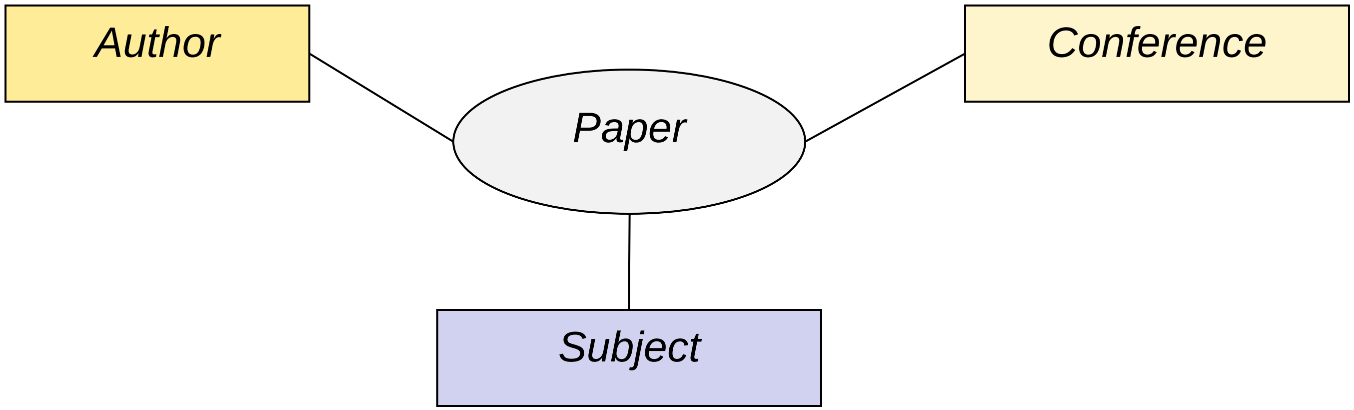
- All the 2,676 conferences and 20,000 authors with most publications, from the time period of year 1998 to year 2007.
- Both conference-author relationships and co-author relationships are used.
- K=15

Table 5: Top-10 Conferences in 5 Clusters Using RankClus

	DB	Network	AI	Theory	IR
1	VLDB	INFOCOM	AAMAS	SODA	SIGIR
2	ICDE	SIGMETRICS	IJCAI	STOC	ACM Multimedia
3	SIGMOD	ICNP	AAAI	FOCS	CIKM
4	KDD	SIGCOMM	Agents	ICALP	TREC
5	ICDM	MOBICOM	AAAI/IAAI	CCC	JCDL
6	EDBT	ICDCS	ECAI	SPAA	CLEF
7	DASFAA	NETWORKING	RoboCup	PODC	WWW
8	PODS	MobiHoc	IAT	CRYPTO	ECDL
9	SSDBM	ISCC	ICMAS	APPROX-RANDOM	ECIR
10	SDM	SenSys	CP	EUROCRYPT	CIVR

Handling Multi-Typed Information Networks

- RankClus works well on bi-typed information networks
- Extension of bi-type network model to star-network model
  - DBLP: Author - paper - conference - title (subject)
  - Netclus model



NetClus: Database System Cluster

database 0.0995511
databases 0.0708818
system 0.0678563
data 0.0214893
query 0.0133316
systems 0.0110413
queries 0.0090603
management 0.00850744
object 0.00837766
relational 0.0081175
processing 0.00745875
based 0.00736599
distributed 0.0068367
xml 0.00664958
oriented 0.00589557
design 0.00527672
web 0.00509167
information 0.0050518
model 0.00499396
efficient 0.00465707

VLDB 0.318495
SIGMOD Conf. 0.313903
ICDE 0.188746
PODS 0.107943
EDBT 0.0436849

author	rank score
Serge Abiteboul	0.0472111
Victor Vianu	0.0348510
Jerome Simeon	0.0324529
Michael J. Carey	0.0288872
Sophie Cluet	0.0282911
Daniela Florescu	0.0241411
Sihem Amer-Yahia	0.0240869
Donald Kossmann	0.0232118
Wenfei Fan	0.0225235
Tova Milo	0.0202201
...	...

Ranking authors in XML

Surajit Chaudhuri 0.00678065
Michael Stonebraker 0.00616469
Michael J. Carey 0.00545769
C. Mohan 0.00528346
David J. DeWitt 0.00491615
Hector Garcia-Molina 0.00453497
H. V. Jagadish 0.00434289
David B. Lomet 0.00397865
Raghu Ramakrishnan 0.0039278
Philip A. Bernstein 0.00376314
Joseph M. Hellerstein 0.00372064
Jeffrey F. Naughton 0.00363698
Yannis E. Ioannidis 0.00359853
Jennifer Widom 0.00351929
Per-Ake Larson 0.00334911
Rakesh Agrawal 0.00328274
Dan Suciu 0.00309047
Michael J. Franklin 0.00304099
Umeshwar Dayal 0.00290143
Abraham Silberschatz 0.00278185

Conclusions

- ASRS data set is rich in text and multidimensional data
- Text cube, comparing cube, and topic cube are interesting new structures and methods for cube space text mining
- To uncover knowledge hidden in massive information networks, we need to integrate text mining, text cube with information network analysis
- Issue: De-identification vs. de-link.
  - Is it possible to get data with de-identification but not de-linked data?
  - Could there be an effective collaborative, distributive mining methodology?
- Much more to be explored on research!

Related Research Publications

- Duo Zhang, Chengxiang Zhai, Jiawei Han, Nikunj C. Oza, Ashok N. Srivastava, "*Topic Cube: Topic Modeling for OLAP on Multidimensional Text Databases*", Journal version invited to special issue of SDM'09
- Cindy Xide Lin, Bolin Ding, Jiawei Han, Nikunj C. Oza, Ashok N. Srivastava, Bo Zhao, Feida Zhu, "*TextCube: Computing IR Measures for Multidimensional Text Database Analysis*", submitted to IEEE Trans. on Knowledge and Data Engineering, 2009
- Duo Zhang, Chengxiang Zhai and Jiawei Han, "*Topic Cube: Topic Modeling for OLAP on Multidimensional Text Databases*Proc. 2009 SIAM Int. Conf. on Data Mining (SDM'09), Sparks, NV, April 2009. (Best of SDM'09)
- Bolin Ding, David Lo, Jiawei Han, and Siau-Cheng Khoo, "Efficient Mining of Closed Repetitive Gapped Subsequences from a Sequence Database", Proc. 2009 Int. Conf. on Data Engineering (ICDE'09), Shanghai, China, Mar. 2009.
- Feida Zhu, Xifeng Yan, Jiawei Han and Philip S. Yu, "*Mining Frequent Approximate Sequential Patterns*", in H. Kargupta, et al., (eds.), Next Generation of Data Mining, Chapman & Hall, 2009.
- Luiz Mendes, Bolin Ding, and Jiawei Han, "*Stream Sequential Pattern Mining with Precise Error Bounds*", Proc. 2008 Int. Conf. on Data Mining (ICDM'08), Pisa, Italy, Dec. 2008.
- Jing Gao, Bolin Ding, Wei Fan, Jiawei Han, and Philip S. Yu, "Classifying Data Streams with Skewed Class Distribution and Concept Drifts", IEEE Internet Computing (Special Issue on Data Stream Management), 12(6): 37-49, 2008
- Chen Chen, Cindy Xide Lin, Xifeng Yan, and Jiawei Han, "On Effective Presentation of Graph Patterns: A Structural Representative Approach", Proc. 2008 ACM Conf. on Information and Knowledge Management (CIKM'08), Napa Valley, CA, Oct. 2008.
- Yintao Yu, Cindy X. Lin, Yizhou Sun, Chen Chen, Jiawei Han, Binbin Liao, Tianyi Wu, ChengXiang Zhai, Duo Zhang, and Bo Zhao, "iNextCube: Information Network-Enhanced Text Cube", Proc. 2009 Int. Conf. on Very Large Data Bases (VLDB'09) (system demo), Lyon, France, Aug. 2009.

Thanks!