National Aeronautics and Space Administration

Scalable Text Data Mining for Improving Aviation Safety

Jiawei Han, ChengXiang Zhai
Department of Computer Science, University of Illinois at Urbana-Champaign
Ashok Srivastava, Nikunj C. Oza
NASA Ames Research Center

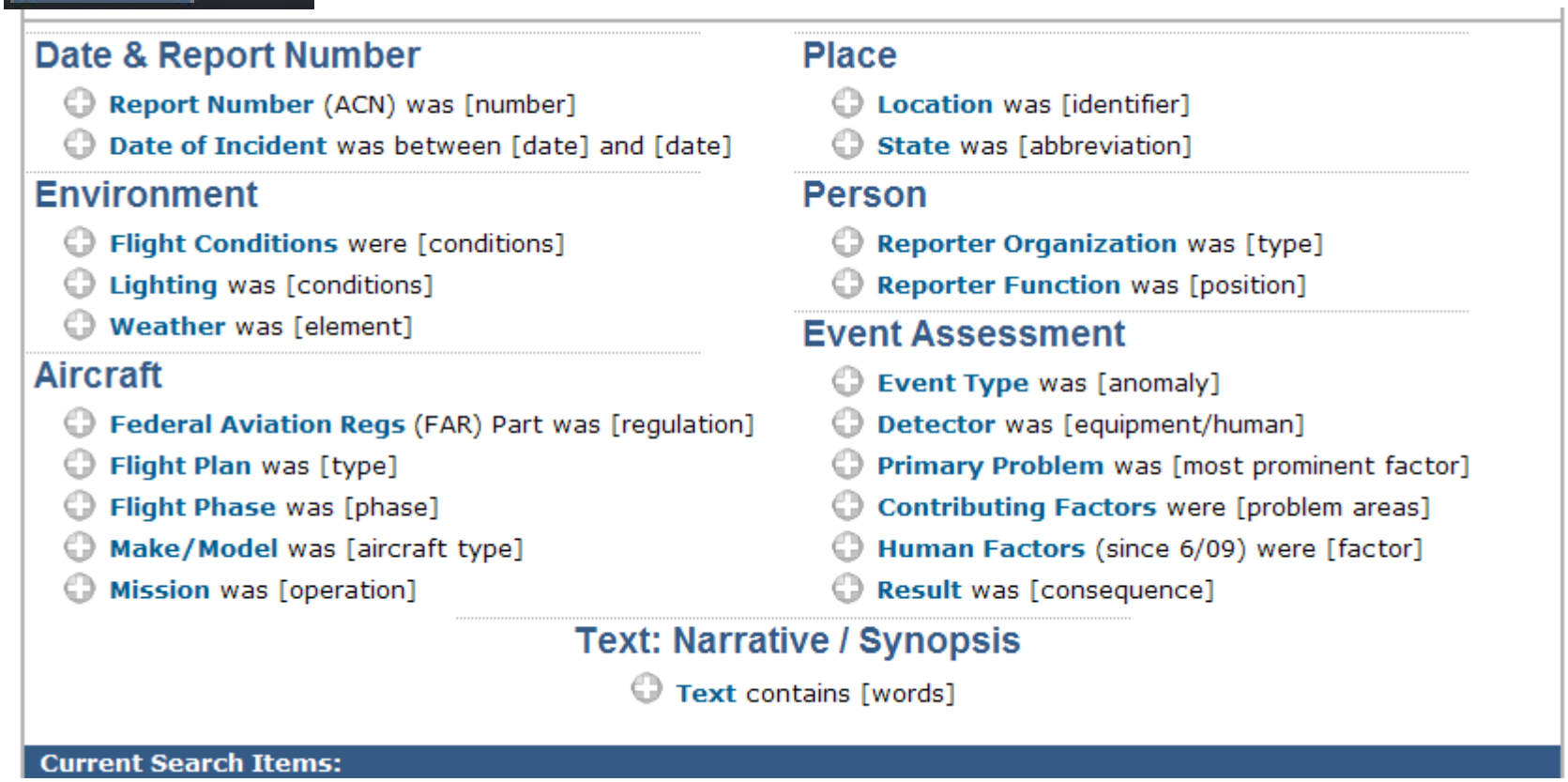
2011 Annual Technical Meeting
May 10–12, 2011
St. Louis, MO

Abundance of text data in the aviation domain

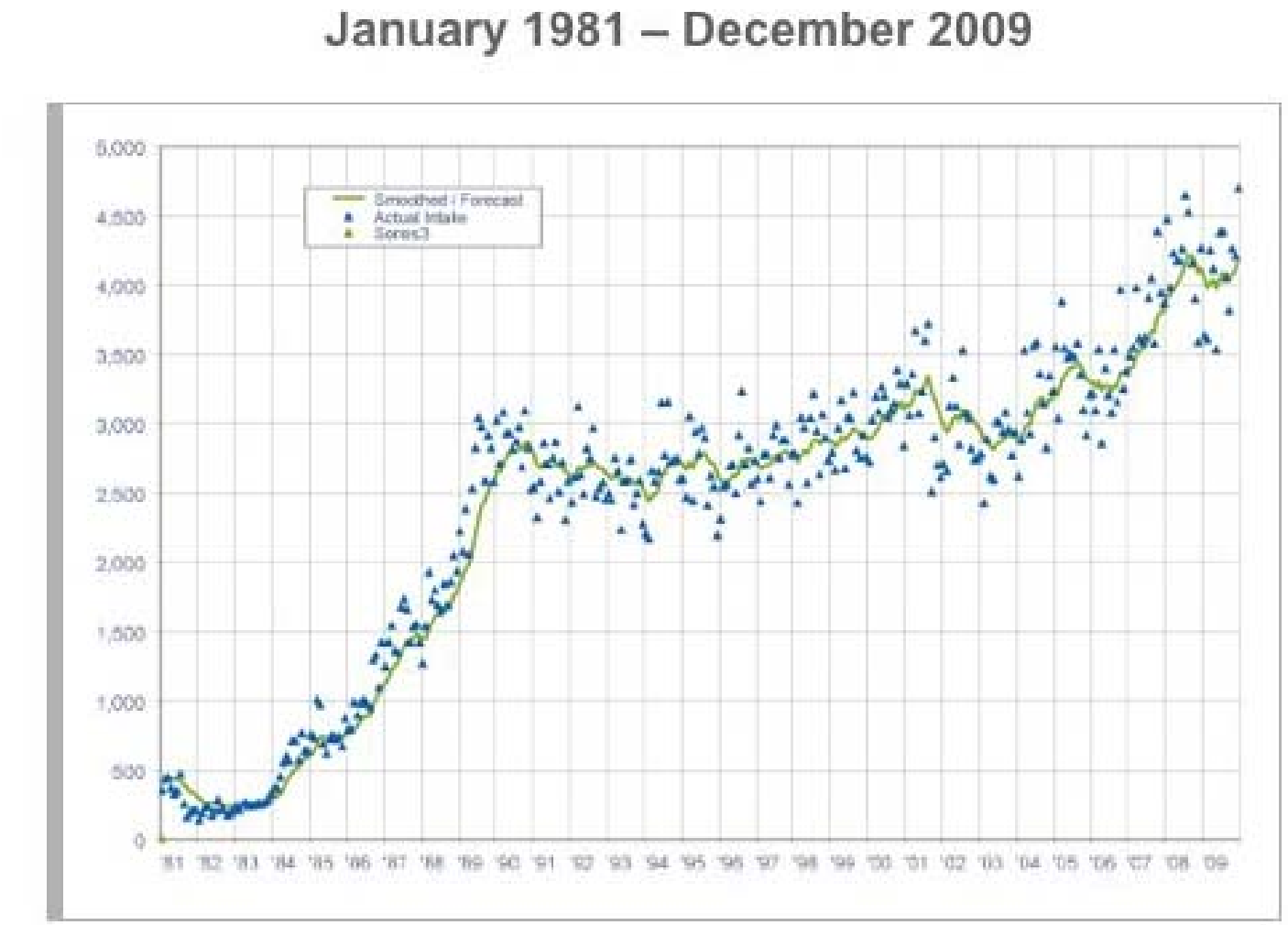
Aviation Safety Reporting System

ASRS

Collecting reports since 1976
>860,000 reports as of Dec. 2009



Monthly intake has been increasing (4k reports/month)



Slide source: http://asrs.arc.nasa.gov/overview/summary.html

Lots of useful knowledge buried in text

ASRS Report ACN: 928983 (Date: 201101, Time: 1801-2400. ...)

We were delayed inbound for about 2 hours and 20 minutes. On the approach there was ice that accumulated on the aircraft. … The Captain wrote up … The flight crew [who picked up the plane] the following morning notified us of an **incorrect remark section write up.** I believe a few years ago, there was a different procedure for writing up aborted takeoffs. I think there was some **confusion as to what the proper write-up for the aborted takeoff was**. A **contributing factor** for this incorrect entry into the log may have been **fatigue**. I had personally been awake for about 14 hours and still had another leg to do. …Also **a contributing factor** is that **this event does not happen regularly….** A **more thorough review and adherence to the operations manual section regarding aircraft status would have prevented this**, [as well as], a better recognition of the onset of **fatigue**. The **manual is sometimes so large that finding pertinent data is difficult**. Even after it was determined that the event had occurred, it took me 15 to 20 minutes to find the section regarding aborted takeoffs.

Challenges

- How can we turn the massive amount of text data into actionable knowledge?

Time	Location	Environment	…	Narrative
199801	TX	Daylight	…	…… I TOLD HIM I WAS AT 2000 FT AND HE SAID OK……
199801	LA	Daylight	…	……WE STOPPED THE DSCNT AT CIRCLING MINIMUMS……
199801	LA	Night	…	……THE TAXI/LNDG LIGHTS VERY DIM. NO OTHER VISIBLE TFC IN SIGHT……
199902	FL	Night	…	……I FEEL WE SHOULD ALL EDUCATE OURSELVES ON CHKLISTS……

How to organize the data to help experts efficiently explore and digest text data?
(e.g. compare the reports before and after a major change in aviation system)
How to help experts analyze a specific type of anomaly in different contexts?
(e.g. what did pilots say about "landing without clearance" at daylight vs. night)

- How can we support an analyst to do this in a general way?
- How can we do this at large scale?

The EventCube Project

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

- **Funded by** NASA IVHM (Integrated Vehicle Health Management)
- **Collaborations of UIUC, UTD, and Boeing**

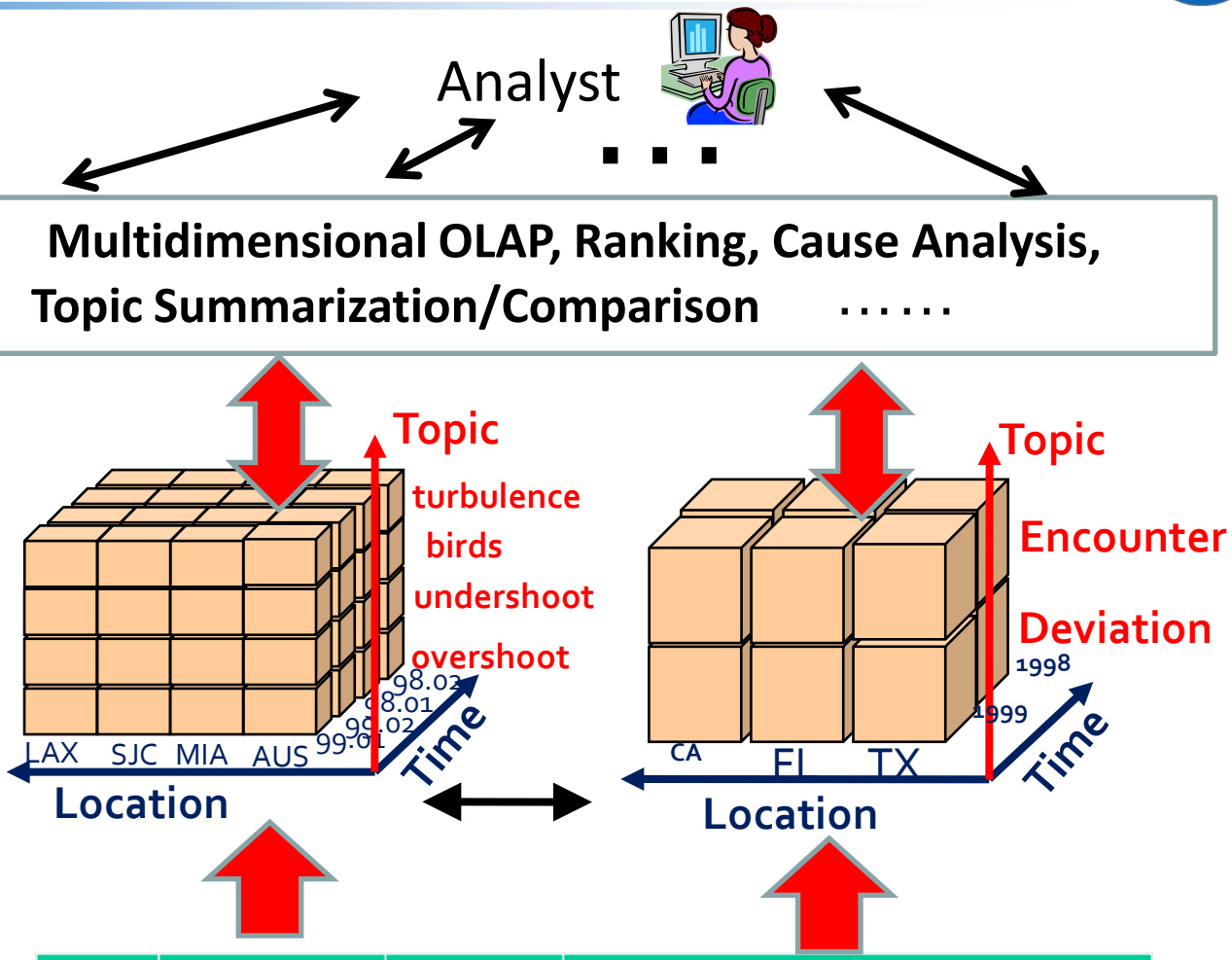
- **Team**
  - **UIUC: Jiawei Han (PI), ChengXiang Zhai**
  - **UTD: Latifur Khan, Vincent Ng, Bhavani Thuraisingham**
  - **Anne Kao (Boeing)**
  - **Graduate students**
- **NASA collaborators: Dr. Ashok Srivastava, Dr. Nikunj C. Oza**

http://eventcube.atwiki.com/

Outline for the rest of the talk

1. **Overview of EventCube**
2. **TopicCube for flexible topic analysis**
3. **Keyword-based mining**
4. **MicroTextCluster for online text summarization**
5. **Demo of iNextCube system**

1. Event Cube: Overview



Light	Anomaly	Phase	Report
Night	equipment : critical	intermediate altitude	...we hit what appeared to be a flock of large white birds. Some of them seem to be ducks..
Daylight	equipment : critical	intermediate altitude	...our flight encountered a small flock of ducks...
Daylight	equipment : less severe	ground: parked	...there were no bird strikes noticed while taxiing..
...	...	...	... ...

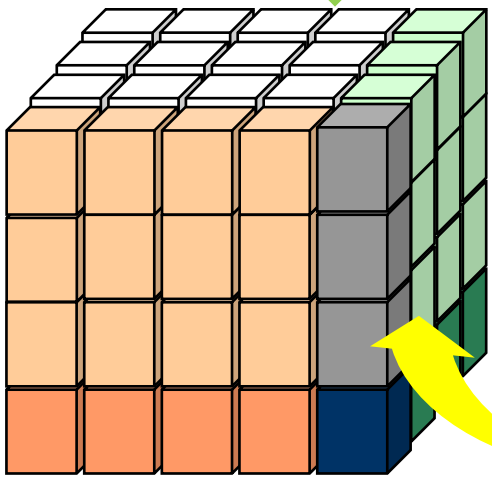
2. Key technique: Text/Topic Cube

- Aggregating text data (ASRS reports) in subspaces
- Heterogeneous: categorical attributes + unstructured text

ACN	Time	Location	Place	Environment	… …	Event	Report
							Text data

- How to combine?
- Our solution:

Cells



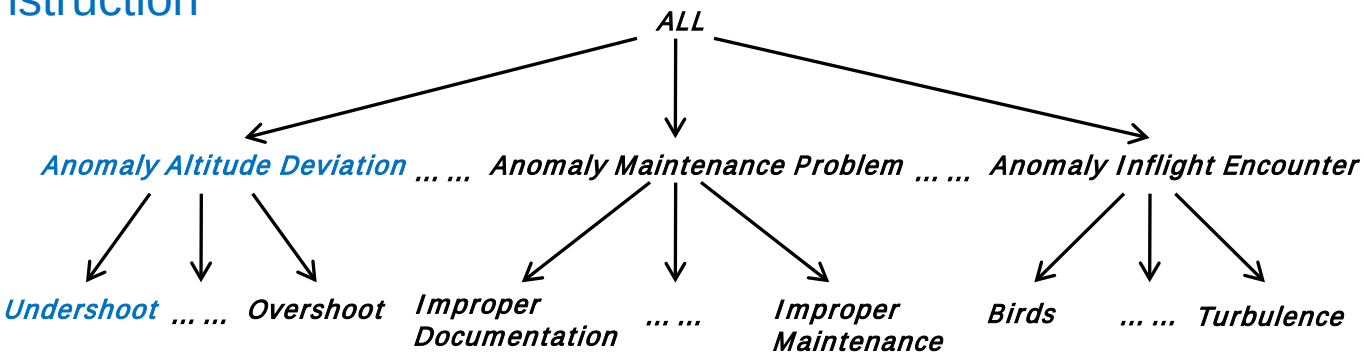
Measure

Text/Topic Model: Unstructured Text

Term/Topic	Weight
altimeter	0.01
leveling	0.008
leveloff	0.007
…	…

Topic Cube Construction

- Construction



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

Topics in daylight

Altitude	0.03
Ft	0.02
Climb	0.01
…	… .

Descent	0.06
Cloud	0.03
Ft	0.01
…	… .

Topics in night

Altitude	0.04
Ft	0.03
Instruct	0.01
…	… .

Descent	0.05
System	0.02
View	0.01
…	… .

Sample Topics in Topic Cube

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

Sample Text:

...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...

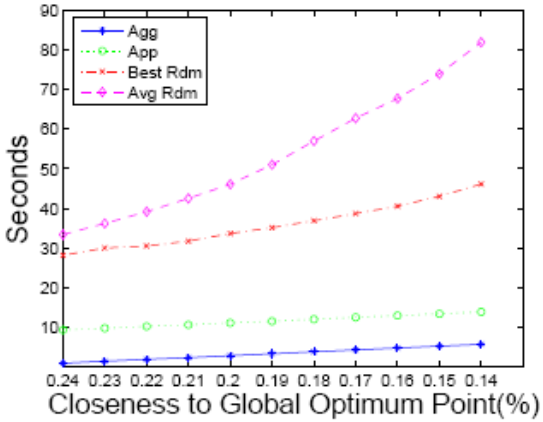
Topic Cube: Efficiency Experiments

Agg: our aggregation method

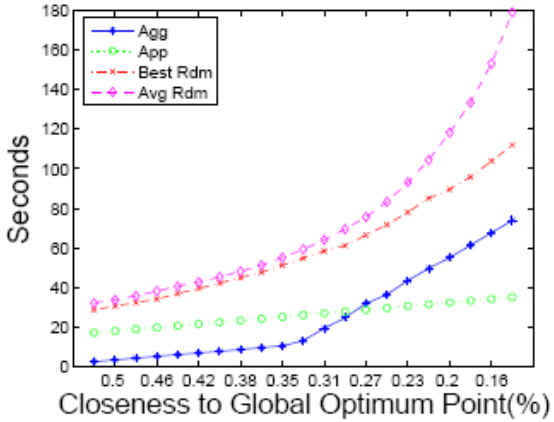
App: Agg with only top *K* words in each topic

Best Rdm: one iteration of EM starting with a random point

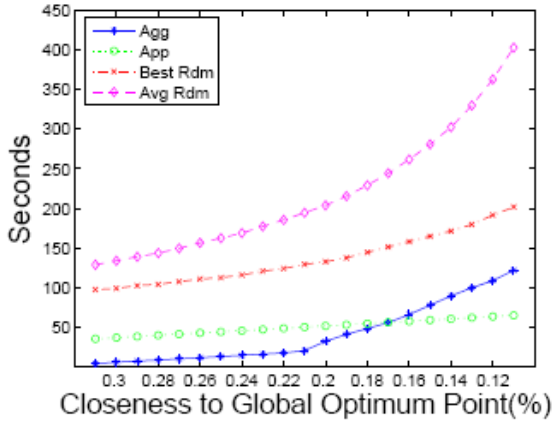
Avg Rdm: average time cost per iteration in standard PLSA



(a) Cell=(1999, CA, *) with 629 documents



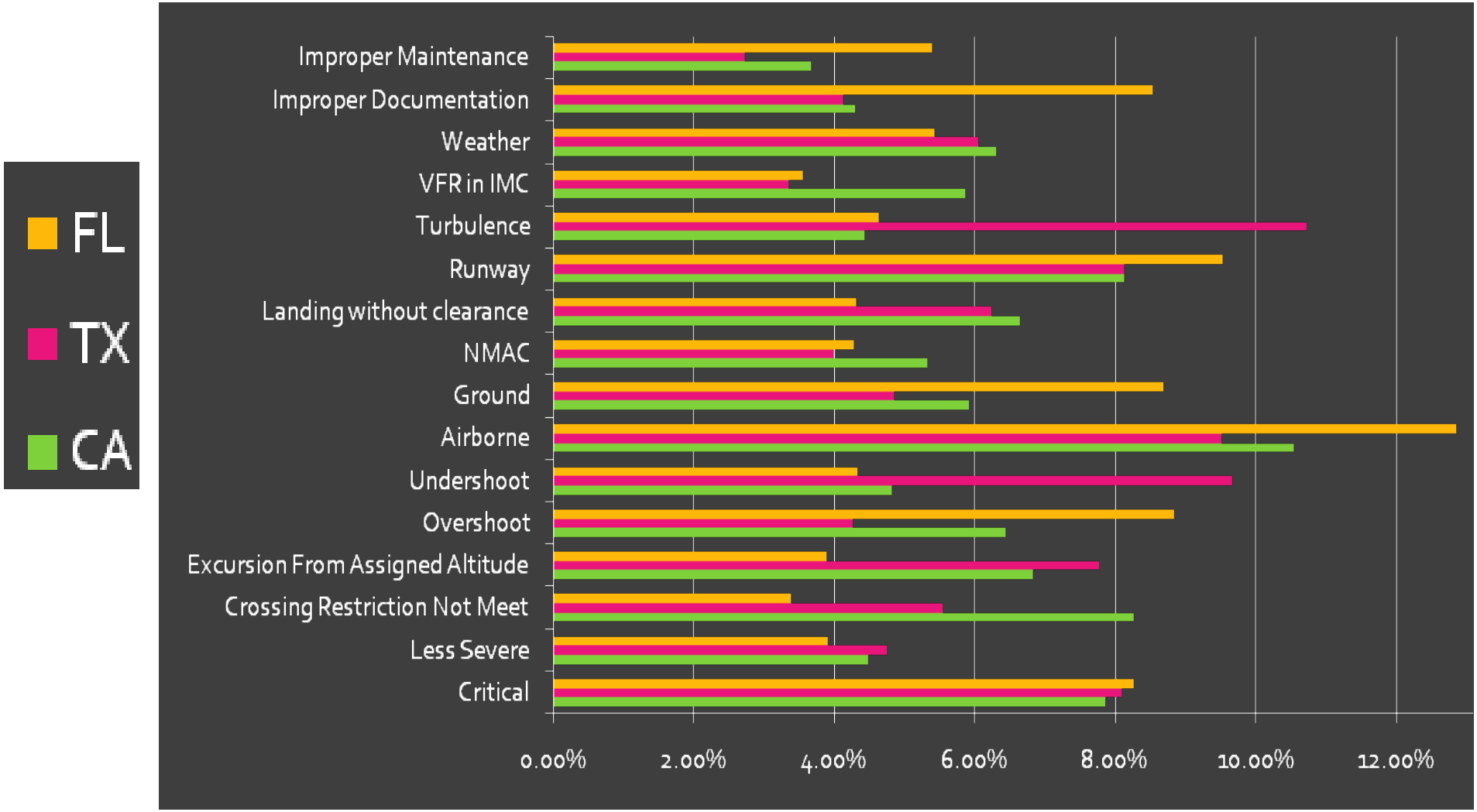
(b) Cell=(1999, *, *) with 1472 documents



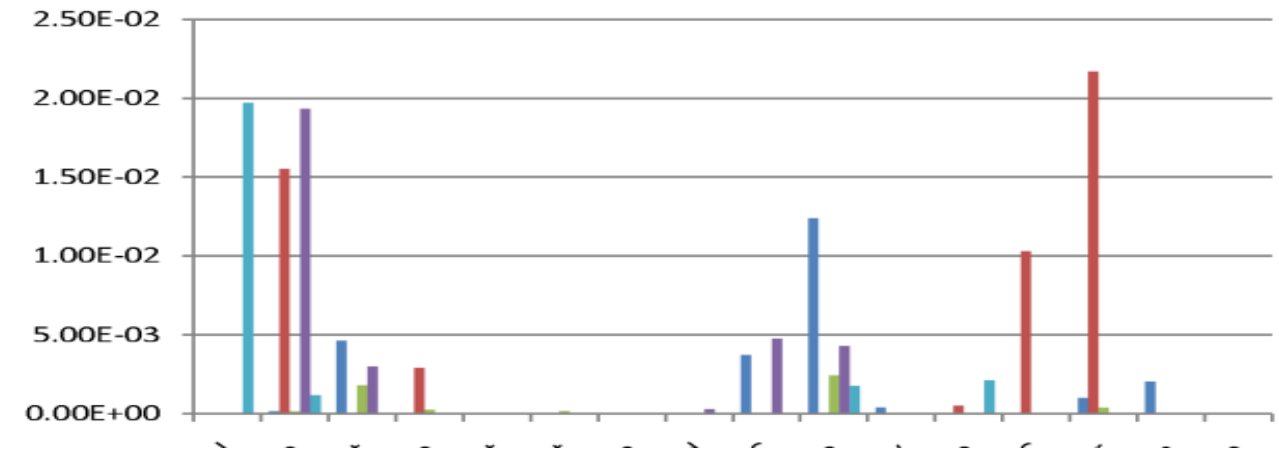
(c) Cell=(*, *, *) with 2733 documents

Sample Topic Coverage Comparison

Comparison of distributions of anomalies in FL, TX, and CA

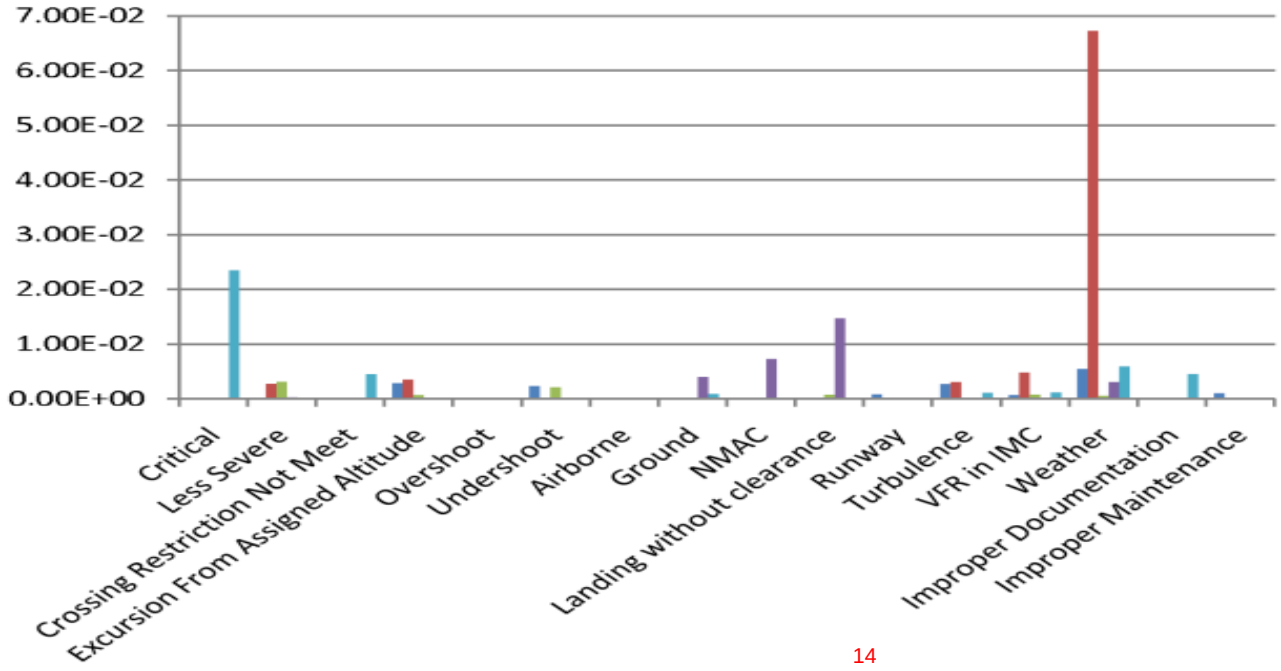


Comparative Analysis of Shaping Factors



Texas

Communication Environment
Physical Environment
Familiarity
Physical Factors
Preoccupation

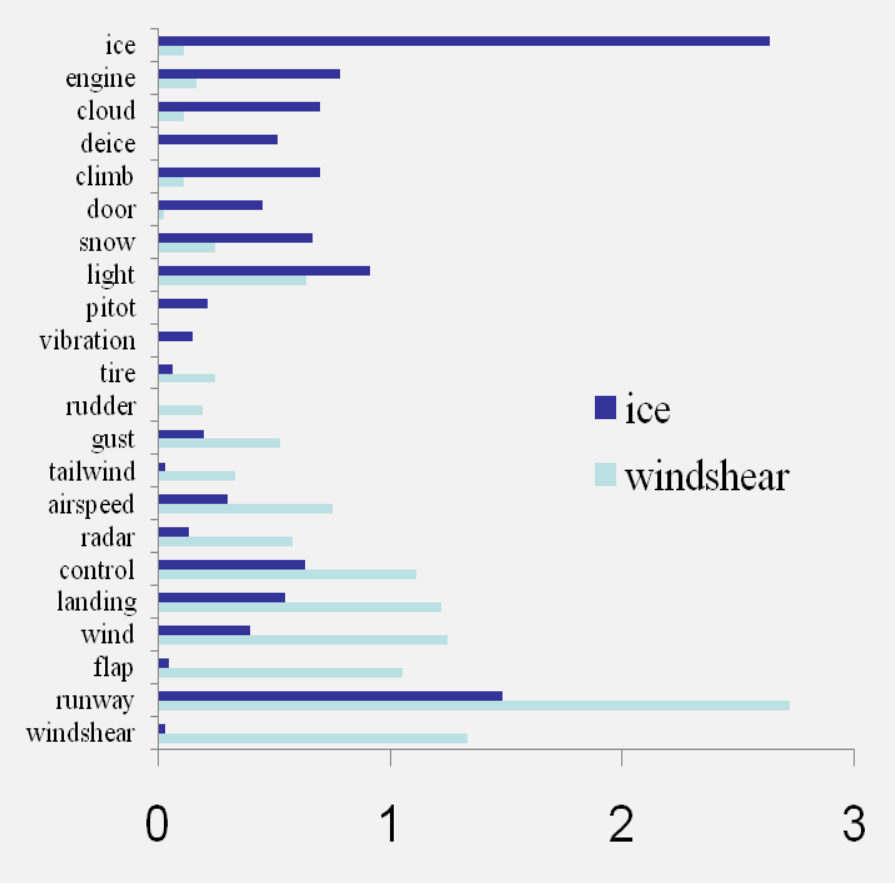


Florida

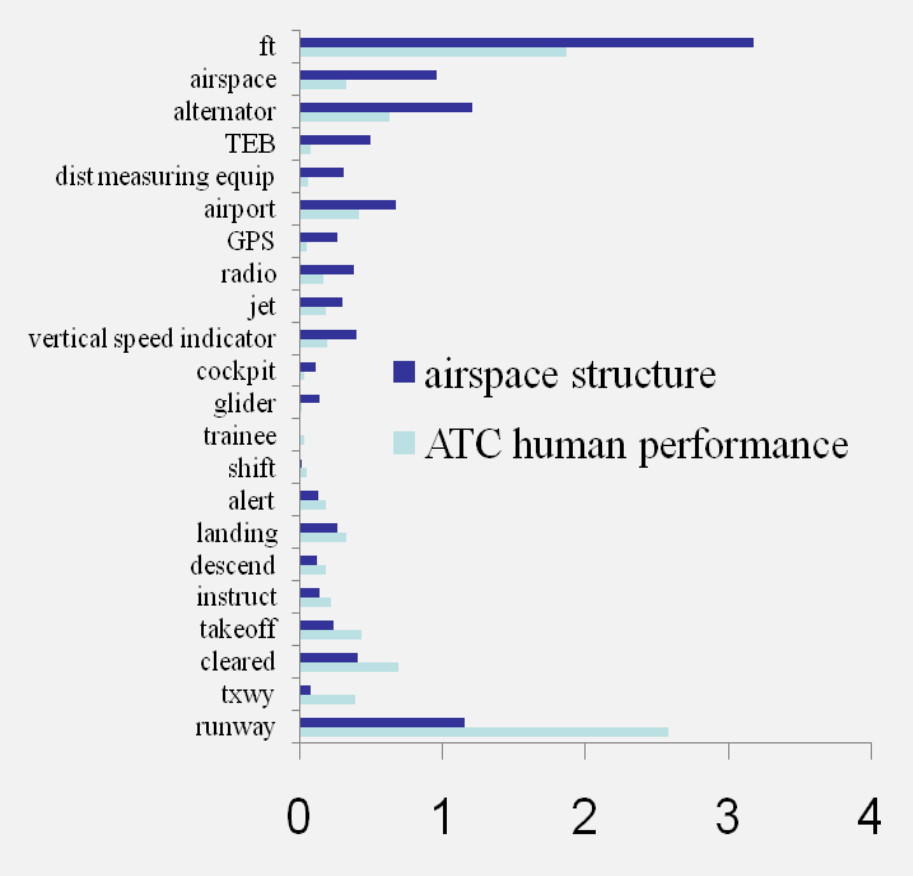
Communication Environment
Physical Environment
Familiarity
Physical Factors
Preoccupation

Text Cube for Comparative Analysis of Sub-Events

"ice" vs. "windshear" in "Environment: Weather Elements"



"airspace structure" vs. "ATC human performance" in "Supplementary: Problem Areas"



Comparison of Average TF of Words

Leverage Sequential Pattern Mining: More Meaningful Units

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

Pattern	Support		
	Anomaly1	Anomaly2	Anomaly3
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

3. Keyword Search

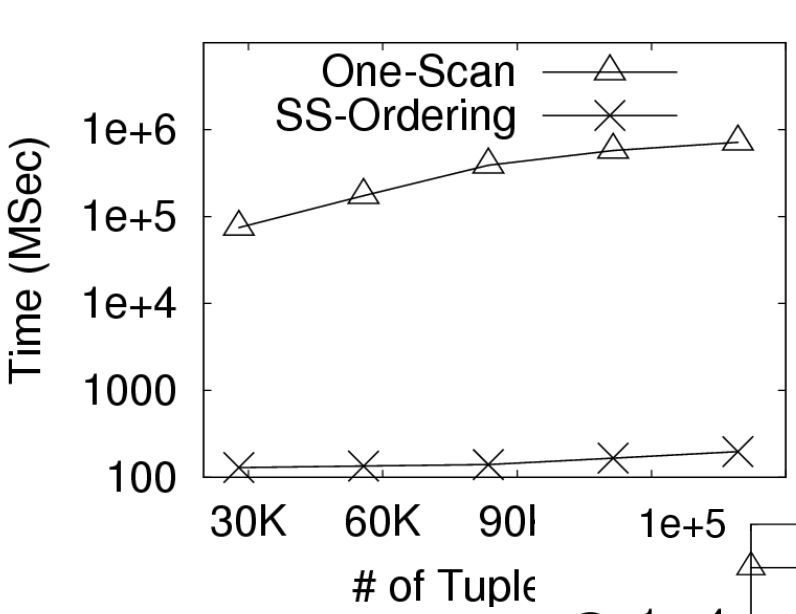
- Find out when evasive action happens in ASRS reports
- Keyword query: “EVASIVE”, “ACTION”

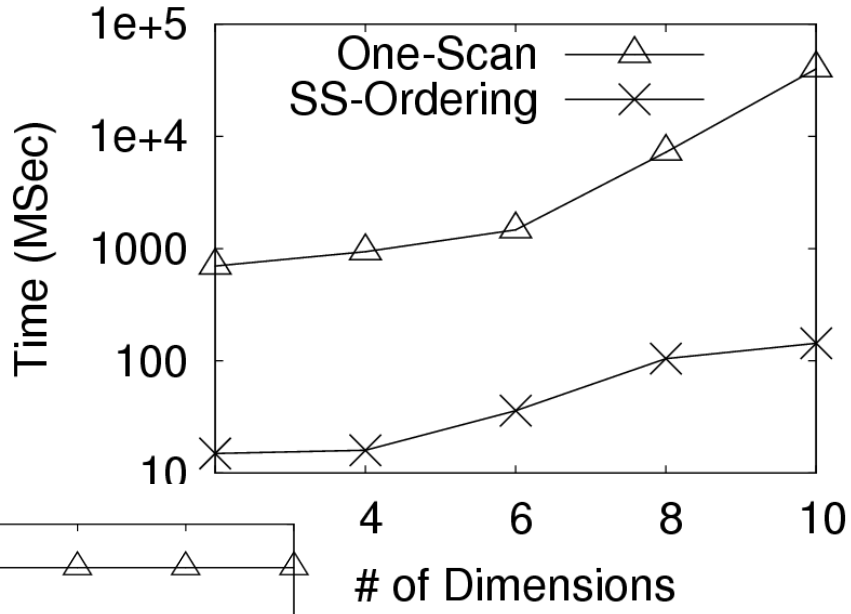
Condition	Light	Phase	Anomaly	Report
VMC	Daylight	descent	conflict	TILL THEY STARTED FLASHING TOO LATE FOR **EVASIVE ACTION** AND SMT X WAS PAST MLT Y
VMC	Daylight	cruise	conflict	I TOOK IMMEDIATE **EVASIVE ACTION** TURNING RIGHT AND DSNDING RAPIDLY TO 1200'.
IMC	Night	descent	equipment	MARCH GCA CTLR TOLD US TO NOTIFY HIM WHEN WE WERE INBND ON THE APCH.

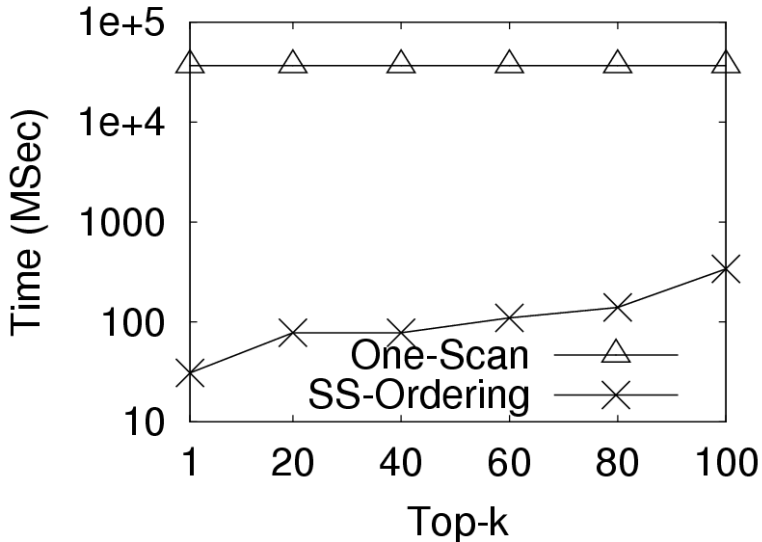
- Main technical challenge: how to score many cells quickly?
- Solutions:
  - Proposed two general strategies: average model, cell document model
  - Proposed search-space-ordering heuristic algorithm to speed up

Experiments: Efficiency (average model)

- Default: # of Docs = 14K; # of Dimensions = 10; k = 80







Sample results of cell ranking

TextCube | TopicCube | **Cell Ranking** | Entity Ranking

Enter key words (space separated) below:

Key Words: rwy excursion

Search

The top ranked cells are:

Rank	Year	State	Person	Weather	Light	Make/Model	Flight Phase	Primary Area	Event Anomaly	Resolutory Action	Score
1	2000	*	*	Rain	Night	*	landing : roll	*	aircraft equipment problem : critical	*	32.9466145601797
2	*	*	*	*	*	McDonnell Douglas	*	Airport	excursion : taxiway	none taken : anomaly accepted	31.6727570181821
3	2000	*	*	*	*	McDonnell Douglas	*	Airport	*	none taken : anomaly accepted	30.8608662261631

4. MiTexCluster Cube for summarization

- **How can we summarize the content in text cells?**
  - **Neutral Summarization**

    Give the most representative documents within a text cell
  - **Topic-biased Summarization**

    Give the most relevant documents to a query within a text cell that also cover the content of the text cell well
- **Example:**

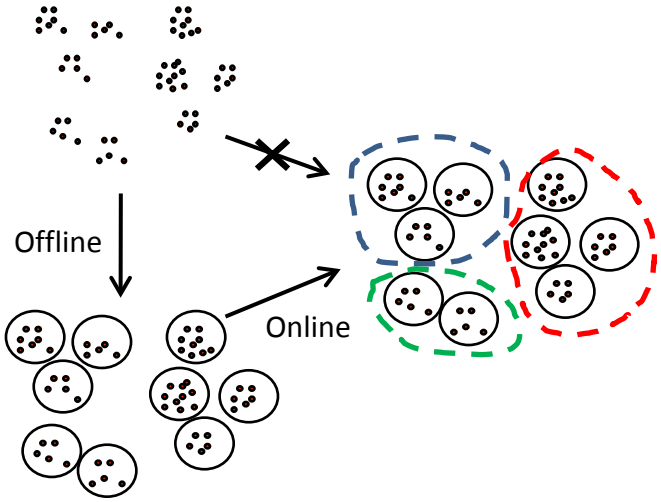
Table 1: An example of text database in ASRS

ACN	Time	Airport	···	Light	Narrative
101285	199901	MSP	···	Daylight	Document 1
101286	199901	CKB	···	Night	Document 2
101291	199902	LAX	···	Dawn	Document 3

  - **What did those reports say about the anomalous events that happened at night in Jan. 1999?**
  - **What did the pilots say about landing anomalies at LAX in 1999?**

Basic idea of MiTexCluster

Goal: Improve online efficiency



Star Schema

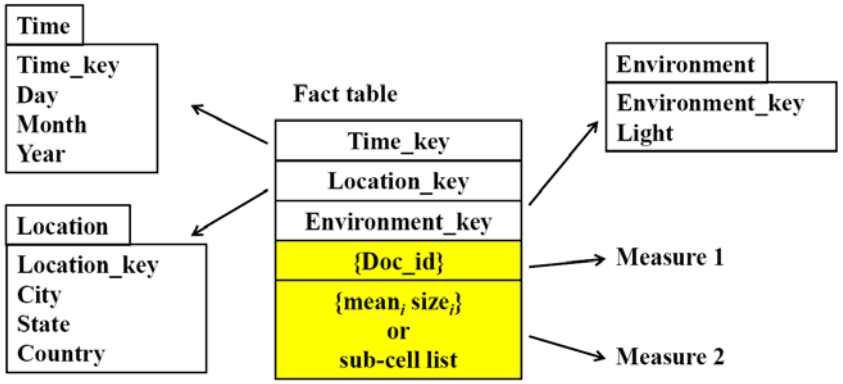


Table 2: An Example of a MiTexCluster Cube

Cell	Doc ID	Content	Micro-Text-Clusters
(Time=1999, Location=TX)	d_1	... due to stronger than forecasted winds and weather going ...	(weather 2.5, wind 1.2, ...), 3
	d_2	... I think that the weather, headwinds, shrinking dew-point/temperature contributed to the fuel emergency ...	
	d_3	... After an hour, the weather had not much improved. We were in the clear for a bit and then hit another cloud bank ...	
	d_4	... so that if we saw the ARPT, we could land ...	(land 2.1, rule 0.9, ...), 2
	d_5	... we were in class G and the IFR rules tell us to land ...	

Sample Results of MiTexCluster

Neutral Summarization

…so that if we saw the ARPT, we could land…

…due to stronger than forecasted winds and weather going…

…resulted in RWY excursion during engine fail…

Summarization biased to "landing"

…so that if we saw the ARPT, we could **land**…

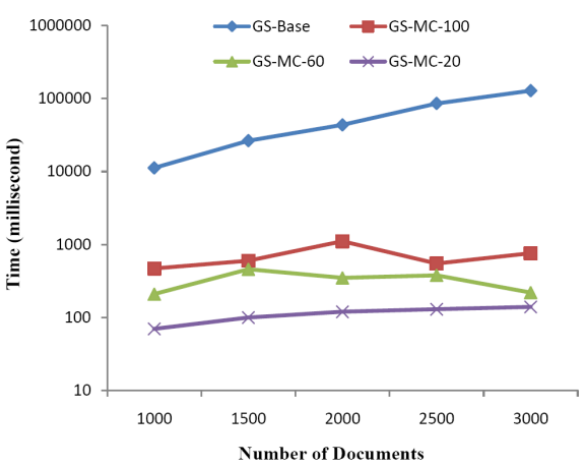
…after an hour, the weather had not much improved which forced us to **land**…

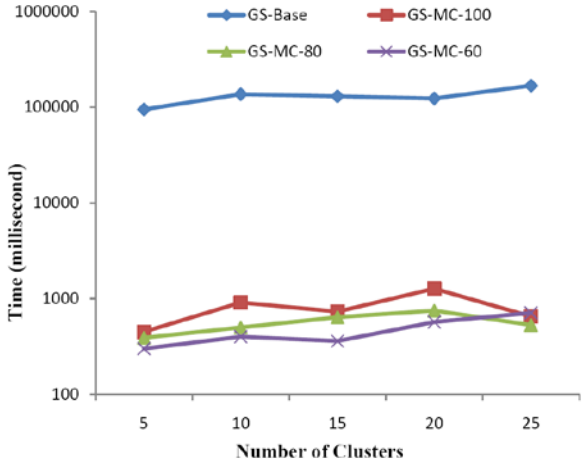
…SMA engine failure, forced **landing** at LGB by instructor…

…we were in class G and the IFR rules tell us to **land**…

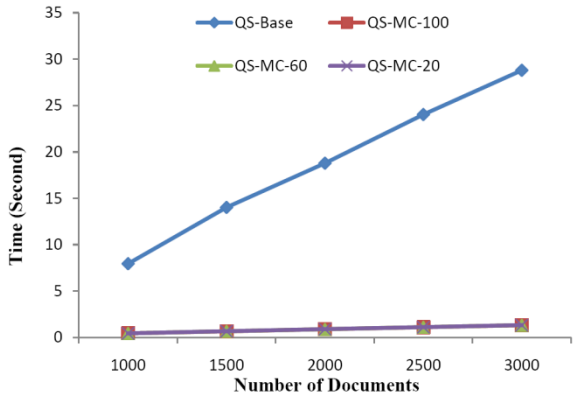
MiTextCluster is More Efficient than Direct Summarization

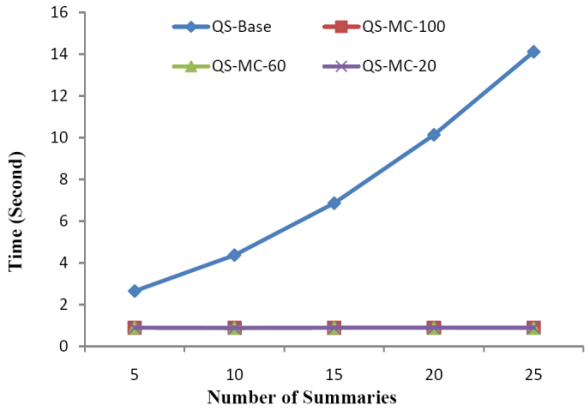
Neutral Summarization: document k-means v.s. micro-cluster k-means





Topic-biased Summarization: MMR v.s. micro-cluster ranking

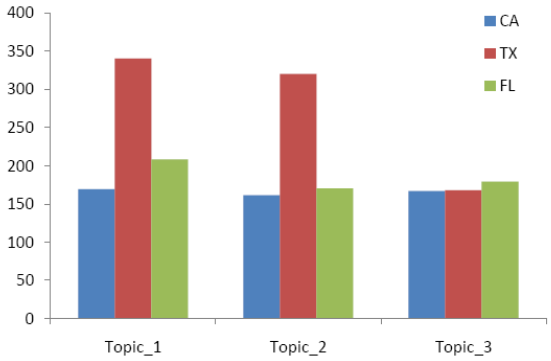




Application: Common Topic Comparison across cells

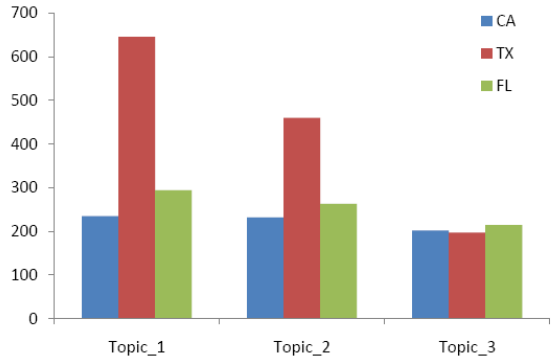
Topic Coverage Comparison across different Cells:

MiTexCluster based results are close to the document based result, and the more micro-clusters stored, the more close to the document based result



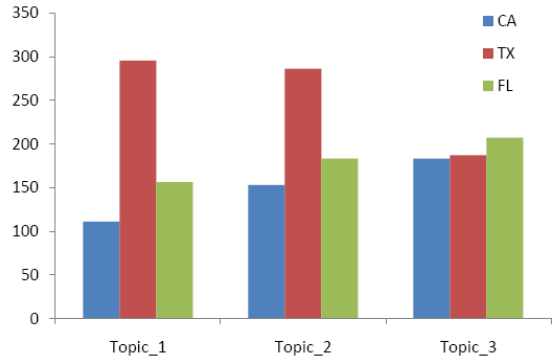
Topic_1: (ft, 2.51808) (tcasii, 2.03836) (deg, 1.98563)
Topic_2: (rwy, 4.20236) (twr, 3.32848) (apch, 2.33843)
Topic_3: (eng, 4.18536) (fuel, 3.06673) (maint, 1.9488)

(a)



Topic_1: (ft, 2.17672) (tfc, 1.41212) (alt, 1.41139)
Topic_2: (rwy, 4.42674) (txwy, 3.18102) (twr, 2.98404)
Topic_3: (eng, 3.58031) (fuel, 2.84583) (rptr, 1.71506)

(b)

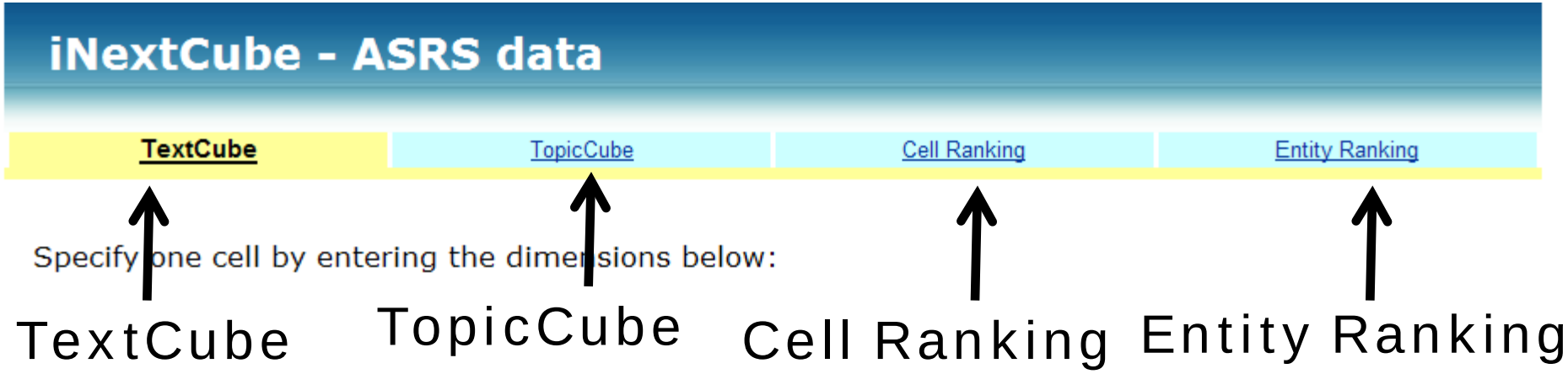


Topic_1: (alt, 2.40233) (ft, 2.34535) (arr, 1.90526)
Topic_2: (rwy, 3.81793) (apch, 2.8462) (twr, 2.64009)
Topic_3: (eng, 3.73188) (fuel, 2.81638) (maint, 1.78779)

(c)

(a) Document based result
(b) MiTexCluster based result (K = 100)
(c) MiTexCluster based result (K = 500)
Y: Number of documents
X: Different locations

5. The iNextCube System



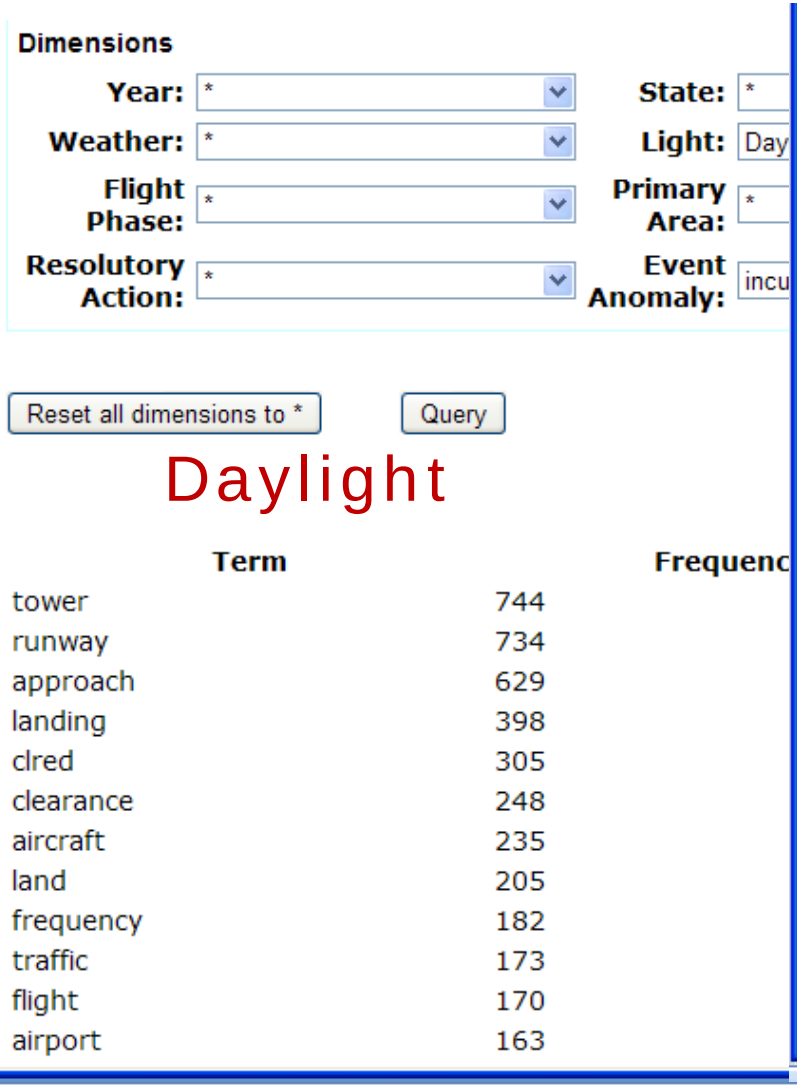
Dimensions

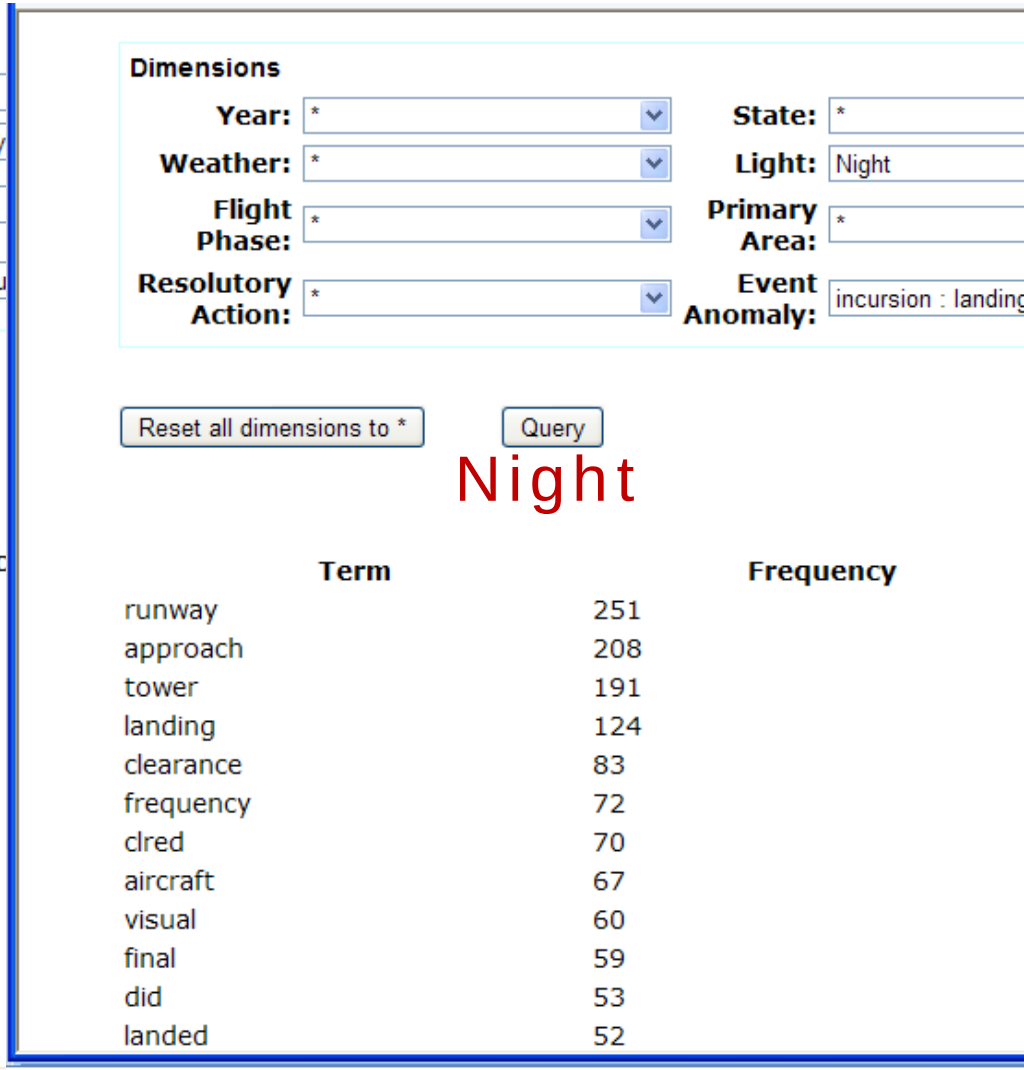
Year:	*	**State:**	*	**Person:**	*
Weather:	*	**Light:**	*	**Make/Model:**	*
Flight Phase:	*	**Primary Area:**	*		
Resolutory Action:	*	**Event Anomaly:**	*		

Reset all dimensions to * Query

http://inextcube.cs.uiuc.edu/nasa/

Sample Results: TextCube





Sample results: TopicCube

Resolutory Action: flight crew : landed in emergen

Less Topics
More Topics

Reset all dimensions to * Query

Landed in emergency

Topic #1

Term	Probability
engine	0.014598
landing	0.013588
fuel	0.012704
aircraft	0.012510
flight	0.011433
zzz	0.009387
gear	0.009353
emergency	0.008537
runway	0.007728
feet	0.006651
approach	0.005716
did	0.005481

Topic #2

Term	Probability
pilot	0.039120
airport	0.019560
did	0.014670
flap	0.014670
visual flight rules	0.012225
flaps	0.012225
landing	0.009780
vfr omni-directional radio range	0.009780
turned	0.009780

Topic #3

Term	Probability
passenger	0.032967
flight	0.032967
aircraft	0.032967
numerous	0.021978
severe	0.021978
turbulence	0.021978
gate	0.021978
got	0.021978
experienced	0.021978
attendants	0.021978
air traffic control	0.021978

Topic #4

Term	Probability
passenger	0.034541
flight	0.022746
medical	0.019377
emergency	0.015164
turbulence	0.010952
aircraft	0.010110
attendant	0.010110
seat	0.009267
zzz	0.009267
captain	0.009267
attendants	0.009267
gate	0.007582

Sample Results: Cell Ranking

iNextCube - ASRS data

TextCube | TopicCube | Cell Ranking | Entity Ranking

Enter key words (space separated) below:

Key Words: fatigue

Search

Fatigue

The top ranked cells are:

Rank	Year	State	Person	Weather	Light	Make/Model	Flight Phase	Primary Area	Event Anomaly	Resolutory Action	Score
1	*	FL	*	*	*	Airbus	*	*	incursion : landing without clearance	*	34.6499617748931
2	*	FL	*	*	Night	*	landing : roll	*	incursion : landing without clearance	none taken : detected after the fact	34.622818450324
3	*	*	*	*	Dawn	*	descent : approach	*	non adherence : far	none taken : anomaly accepted	34.5521442932267
4	2006	FL	*	*	Night	Airbus	*	*	*	*	34.5034423658916
5	2006	*	flight crew : first officer	*	Night	Airbus	*	*	*	none taken : detected after the fact	34.1260320579751
6	*	FL	*	*	Night	*	landing : roll	*	incursion : landing without clearance	*	34.0639315928729
7	*	*	*	*	*	*	*	Cabin Crew Human Performance	non adherence : far	none taken : anomaly accepted	33.9984240321173
8	*	*	*	*	*	*	*	Environmental	excursion : ramp	none taken :	33.9404062436666

Sample results: Entity Ranking

Please choose one anomaly type:

8.1 incursion : landing without clear

Search

Insursion: landing without clearance

The top ranked entities are:

Person

Rank	Entity
1	flight crew : captain
2	flight crew : first officer
3	flight crew : single pilot
4	instruction : instructor
5	instruction : trainee
6	flight crew : relief pilot
7	flight crew : second officer
8	controller : local
9	observation : observer

Weather

Rank	Entity
1	Turbulence
2	Fog
3	Rain
4	Snow
5	Ice
6	Thunderstorm
7	Windshear

Make/Model

Rank	Entity
1	Boeing
2	Cessna
3	Bombardier
4	McDonnell Douglas
5	Piper
6	Airbus
7	Beechcraft
8	Embraer
9	British Aerospace
10	Mooney
11	Fairchild Dornier
12	ATR
13	Saab
14	Lockheed

Flight Phase

Rank	Entity
1	landing : roll
2	descent : approach
3	descent : vacating altitude
4	ground : taxi
5	landing : touch and go
6	cruise : level
7	climbout : takeoff
8	descent : intermediate altitude
9	ground : takeoff roll

Primary Area

Rank	Entity
1	Flight Crew Human Performance
2	ATC Human Performance
3	Ambiguous
4	Environmental Factor
5	Company
6	Weather
7	ATC Facility
8	Navigational Facility
9	Passenger Human Performance
10	Aircraft
	Maintenance

Resolutory Action

Rank	Entity
1	none taken : detected after the fact
2	none taken : anomaly accepted
3	controller : issued advisory
4	flight crew : landed as precaution
5	controller : issued new clearance
6	none taken : insufficient time
7	none taken :

Sample results of iNextCube: Entity Ranking

Please choose one anomaly type:

2 airspace violation

Search

Airspace violation

The top ranked entities are:

Person

Rank	Entity
1	flight crew : single pilot
2	flight crew : captain
3	controller : radar
4	instruction : instructor
5	flight crew : first officer
6	controller : approach
7	instruction : trainee
8	controller : departure
9	controller :

Weather

Rank	Entity
1	Turbulence
2	Thunderstorm
3	Rain
4	Fog
5	Snow
6	Windshear
7	Ice

Make/Model

Rank	Entity
1	Cessna
2	Piper
3	Beechcraft
4	Boeing
5	Bombardier
6	Mooney
7	McDonnell Douglas
8	Airbus
9	Bell Helicopter
10	Robinson Helicopter Company
11	Embraer
12	Dassault

Flight Phase

Rank	Entity
1	cruise : level
2	descent : approach
3	climbout : intermediate altitude
4	descent : intermediate altitude
5	climbout : initial
6	climbout : takeoff
7	cruise : enroute altitude change

Primary Area

Rank	Entity
1	Flight Crew Human Performance
2	ATC Human Performance
3	FAA
4	Ambiguous
5	Airspace Structure
6	Weather
7	Chart Or Publication
8	ATC Facility
9	Aircraft
10	Company
11	Environmental

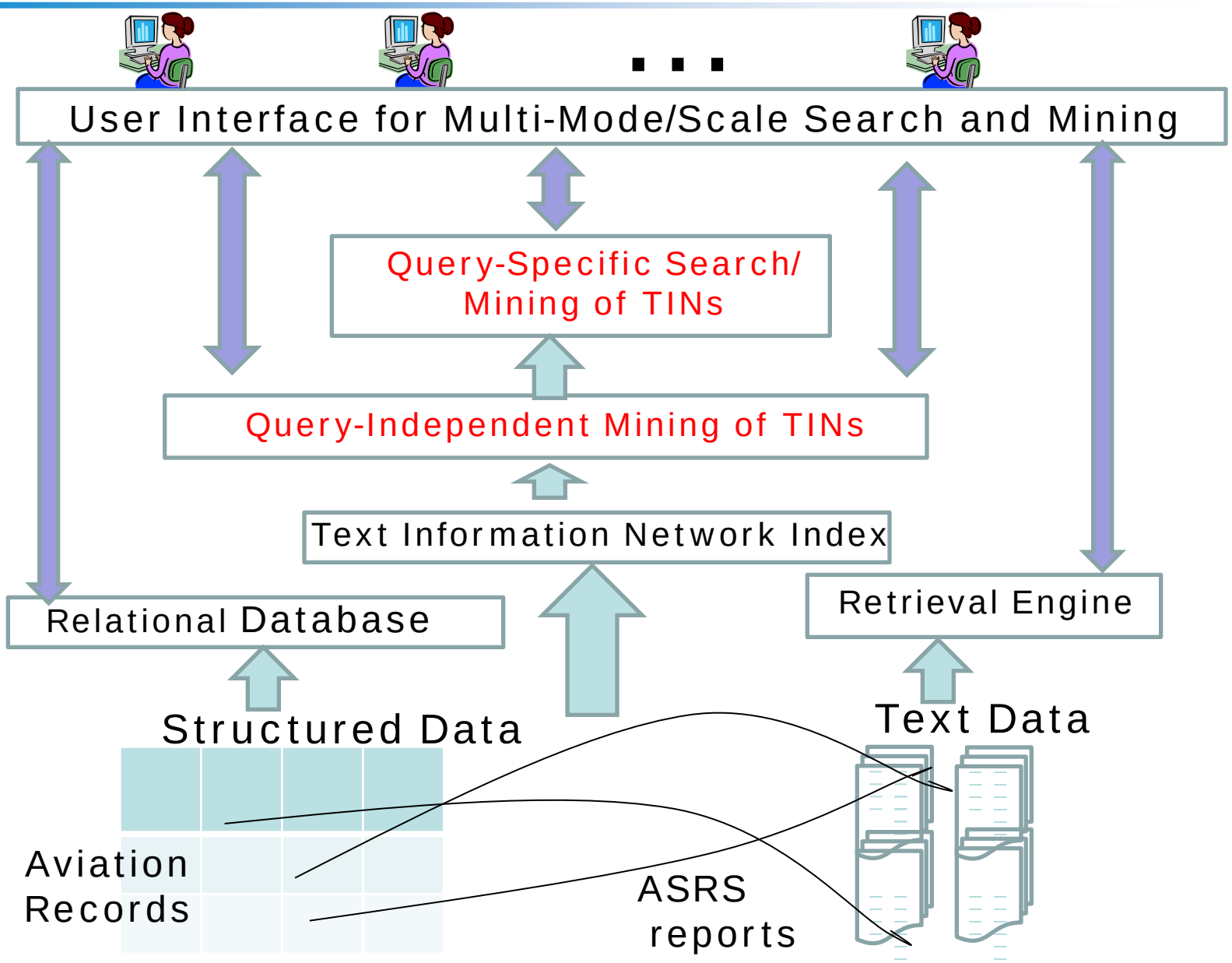
Resolutory Action

Rank	Entity
1	none taken : detected after the fact
2	controller : issued advisory
3	controller : issued new clearance
4	flight crew : exited penetrated airspace
5	none taken : anomaly accepted
6	controller : issued alert

Summary

- TextCube/TopicCube provides a general and scalable support for analyzing topics in text data in high-dimensional databases
- Cell ranking and entity ranking enable flexible mining of topical cells and entities
- MiTextCluster enables efficient online summarization
- iNextCube system supports multiple ways to mine and analyze ASRS reports
- Multiple mining applications for improving aviation safety can be potentially supported with these component technologies

Future Work 1: Large-Scale Integrative Text Mining



Future Work 2: Text Mining for Proactive Prevention of Aviation Incidents

- Semantic Analysis of Text (Information Extraction):
  - How to recognize entities (e.g., people, devices, time, location) ?
  - How to recognize relations (e.g., what happened at what time)?
  - How to recognize sentences of special semantic categories (e.g., contributing factors, suggestions)?
- Mining ASRS to discover knowledge for preventing incidents
  - What problems and causal factors are **increasingly** reported in ASRS?
  - What **suggestions** have been made by reporters in ASRS?
  - How can we discover knowledge about human factors?
- Combine features extracted from text with other non-textual features to **improve statistical prediction models**

References

- **Duo Zhang, ChengXiang Zhai, Jiawei Han, Ashok Srivastava, Nikunj Oza. Topic Modeling for OLAP on Multidimensional Text Databases: Topic Cube and its Applications, *Statistical Analysis and Data Mining*, Vol. 2, pp.378-395, Special Issue on the Best of SDM'09.**
- 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.
- Bolin Ding, Bo Zhao, Cindy Xide Lin, Jiawei Han, Chengxiang Zhai, Ashok Srivastava, Nikunj C. Oza, "Efficient Keyword-Based Search for Top-K Cells in Text Cube", *IEEE Transactions on Knowledge and Data Engineering (TKDE)* (Special Issue: Keyword Search on Structured Data), accepted, Dec. 2010.
- 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.

More publications can be found in Dashlink:
https://c3.ndc.nasa.gov/dashlink/

Acknowledgments

- EventCube Project is funded by NASA IVHM (Integrated Vehicle Health Management) Program
- Graduate students at UIUC: Duo Zhang, Bolin Ding, Cindy Xide Lin, Yintao Yu, Bo Zhao
- Other collaborators at UT Dallas & Boeing

Thank You!