Incremental Discovery of Prominent Situational Fact

30th IEEE International Conference on Data Engineering, Chicago, IL, 2014

Afroza Sultana[1], Naeemul Hassan[1], Chengkai Li[1], Jun Yang[2], Cong Yu[3]

[1]University of Texas at Arlington, [2]Duke University, [3]Google Research

Situational Facts

- Sports: "Paul George had 21 points, 11 rebounds and 5 assists to become the first Pacers player with a 20/10/5 (points/rebounds/assists) game against the Bulls since Detlef Schrempf in December 1992."
- Social Media: "The social world's most viral photo ever generated 3.5 million likes, 170,000 comments and 460,000 shares by Wednesday afternoon."
- Stock Data: Stock A becomes the first stock in history with price over $300 and market cap over $400 billion.
- Weather Data: Today's measures of wind speed and humidity are x and y, respectively. City B has never encountered such high wind speed and humidity in March.
- Criminal Records: There were 50 DUI arrests and 20 collisions in city C yesterday, the first time in 2013.

A Mini-world of Basketball Gamelogs

id	player	day	month	season	team	opp_team	pts	ast	reb
t_1	Bogues	11	Feb.	1991-92	Hornets	Hawks	4	12	5
t_2	Seikaly	13	Feb.	1991-92	Heat	Hawks	24	5	15
t_3	Sherman	7	Dec.	1993-94	Celtics	Nets	13	13	5
t_4	Wesley	4	Feb.	1994-95	Celtics	Nets	2	5	2
t_5	Wesley	5	Feb.	1994-95	Celtics	Timberwolves	3	5	3
t_6	Strictland	3	Jan.	1995-96	Blazers	Celtics	27	18	8
t_7	Wesley	25	Feb.	1995-96	Celtics	Nets	12	13	5

Last tuple appended to table

Wesley had 12 points, 13 assists and 5 rebounds on February 25, 1996 to become the first player with a 12/13/5 (points/assists/rebounds) in February.

Wesley had 13 assists and 5 rebounds on February 25, 1996 to become the second Celtics player with a 13/5 (assists/rebounds) game against the Nets.

Problem Definition

- Constraint (C): $d_1=v_1 \wedge d_2=v_2 \wedge \ldots \wedge d_n=v_n, v_i \in dom(d_i) \cup \{*\}$
  - team=*Celtics* ∧ opp_team=*Nets*
- Constraint-Measure Pair (C, M): Combination of a constraint and measure subspace
  - (team=*Celtics* ∧ opp_team=*Nets*,{assists,rebounds})
- Contextual skyline: skyline regarding (C, M)
  - $\sigma_{\text{team=Celtics} \wedge \text{opp\_team=Nets}}(R)$, M={assists,rebounds}
    - $\{t_3\}$

Tuples capturing real world events appended to table

id	player	day	month	season	team	opp_team	pts	ast	reb
t_1	Bogues	11	Feb.	1991-92	Hornets	Hawks	4	12	5
t_2	Seikaly	13	Feb.	1991-92	Heat	Hawks	24	5	15
t_3	Sherman	7	Dec.	1993-94	Celtics	Nets	13	13	5
t_4	Wesley	4	Feb.	1994-95	Celtics	Nets	2	5	2
t_5	Wesley	5	Feb.	1994-95	Celtics	Timberwolves	3	5	3
t_6	Strictland	3	Jan.	1995-96	Blazers	Celtics	27	18	8
t_7	Wesley	25	Feb.	1995-96	Celtics	Nets	12	13	5

Find constraint-measure pair (C,M) such that t is in the contextual skyline.

Constraint	Measure
month=*Feb*	pts, ast, rb
opp_team=*Nets*	ast, rb
team=*Celtics* ∧ opp_team=*Nets*	ast, rb
…	…

Template

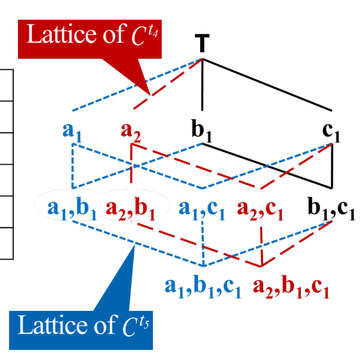
Wesley had 12 points, 13 assists and 5 rebounds on February 25, 1996 to become the first player with a 12/13/5 (points/assists/rebounds) in February.

Related Work

- Conventional skyline analysis (Borzsonyi et al. ICDE 2001)
  - Given question, find answer
- Compressed Skycube (Xia et al. SIGMOD 2006)
  - Update compressed skycube in monitoring fashion
- Prominent Analysis by Ranking (Wu et. Al. VLDB 2009)
  - Static data, onetime query
  - Find the contexts where an object is ranked high in a single scoring attribute

Modeling

id	d_1	d_2	d_3	m_1	m_2
t_1	a_1	b_2	c_2	10	15
t_2	a_1	b_1	c_1	15	10
t_3	a_2	b_1	c_2	17	17
t_4	a_2	b_1	c_1	20	20
t_5	a_1	b_1	c_1	11	15



Tuple Satisfied Constraint C^t: If $\forall d_i \in \mathcal{D}$, $C.d_i=*$ or $C.d_i=t.d_i$, t satisfies C.

Lattice Intersection: $C^{t_4,t_5}=C^{t_4} \cap C^{t_5}$

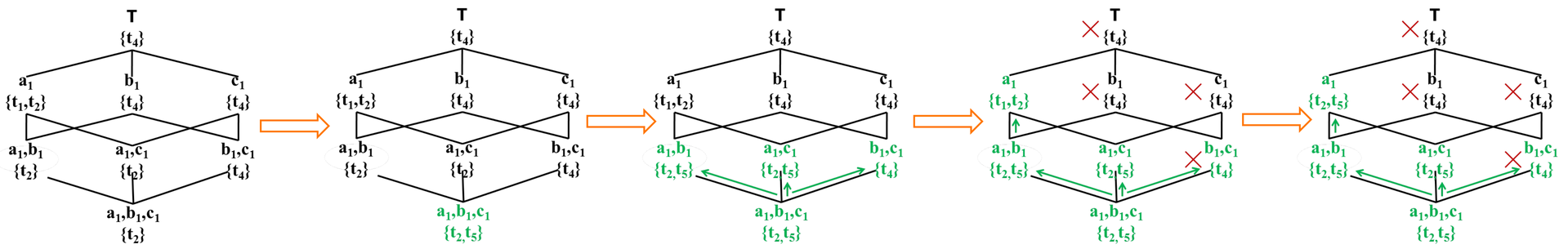
Challenges and Ideas

- Exhaustive comparison with every tuple
  - ✓Tuple reduction
    - $t_4 \succ_{\{m_1,m_2\}} t_3 \succ_{\{m_1,m_2\}} t_5 \Rightarrow t_4 \succ_{\{m_1,m_2\}} t_5$
    - Comparison with skyline tuples are enough
- Under every constraint
  - ✓Constraint pruning
    - In $C^{t,t'}$, one comparison on t and t' is enough
- Over every measure subspace
  - ✓Sharing computation across measure subspaces
    - Reusing computations on full space in subspaces

Algorithm BottomUp

- Stores a tuple for every such constraint that qualifies it as a contextual skyline tuple
- Traverses the constraints in C^t in a bottom-up, breadth-first manner

Total 6 comparisons in this scenario

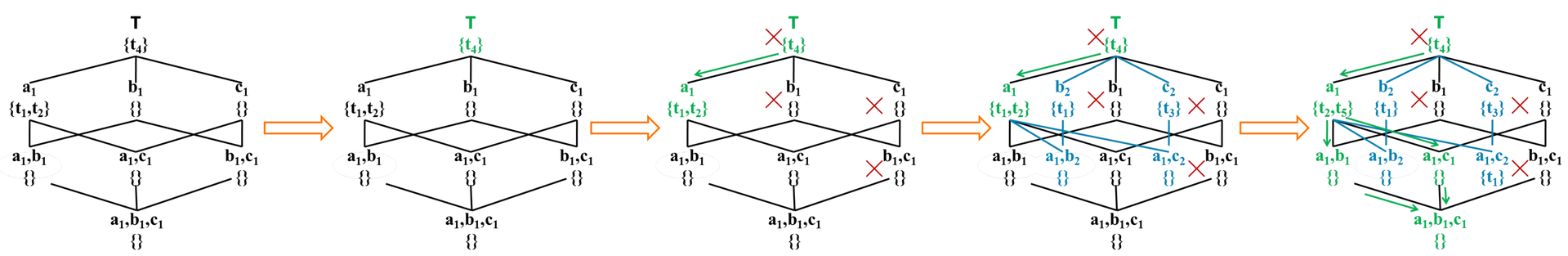


Algorithm TopDown

Skyline Constraints: Constraints whose contextual skylines include t.

Maximal Skyline Constraints: Constraints not subsumed by any other skyline constraints of t.

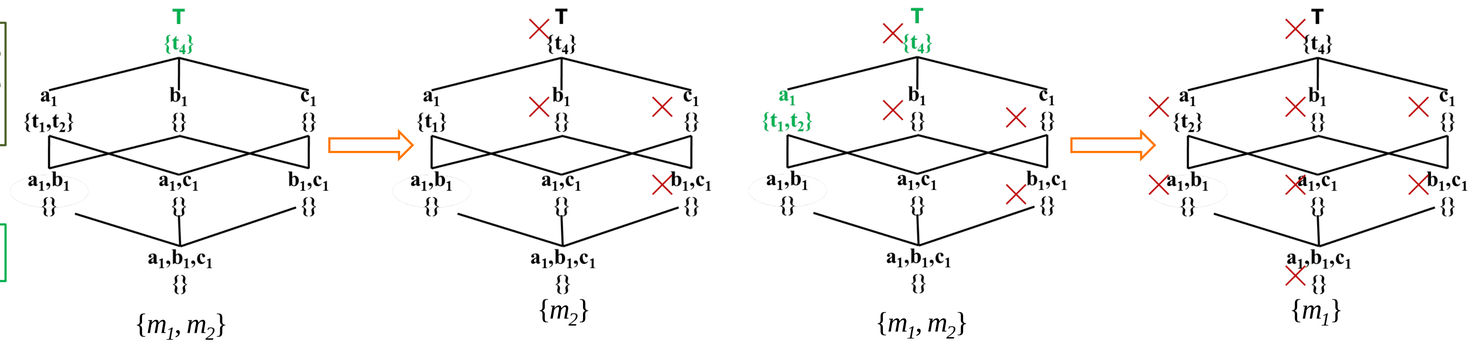
Total 3 comparisons in this scenario



Algorithm STopDown

Computation over full space is enough in finding skyline constraints in subspaces.

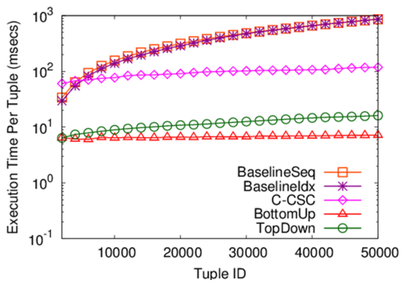
Skips 3 comparisons



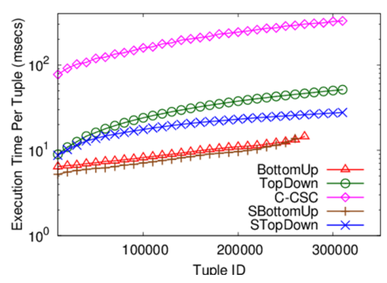
Experiment Setup

- NBA Dataset
  - 317,371 tuples of NBA box scores from 1991-2004 seasons
  - 8 dimension and 7 measure attributes
- Weather Dataset
  - 7.8 million tuples of weather forecast from different locations of six countries & regions of UK
  - 7 dimension and 7 measure attributes

Memory-Based Implementation

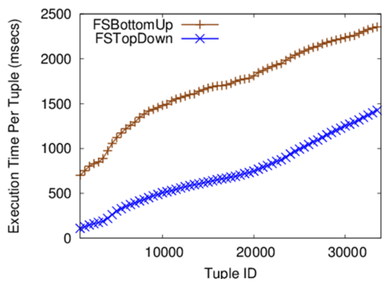


NBA Dataset

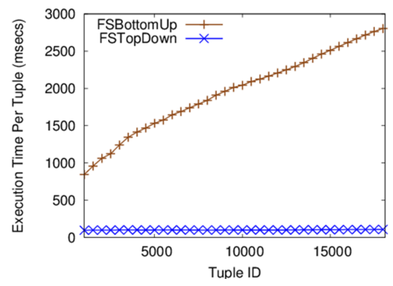


NBA Dataset

File-Based Implementation



NBA Dataset



Weather Dataset

Discovered Facts

- Lamar Odom had 30 points, 19 rebounds and 11 assists on March 6, 2004. No one before had a better or equal performance in NBA history.
- Allen Iverson had 38 points and 16 assists on April 14, 2004 to become the first player with a 38/16 (points/assists) game in the 2004-2005 season.
- Damon Stoudamire scored 54 points on January 14, 2005. It is the highest score in history made by any Trail Blazers.