

Generating Preview Tables for Entity Graphs

#Ning Yan, *Sona Hasani, *Abolfazl Asudeh, *Chengkai Li

#Huawei U.S. R&D Center

*University of Texas at Arlington, Innovative Database and Information Systems Research (IDIR) Laboratory

SIGMOD 2016, June 30th, 2016



Ultra-heterogeneous Entity Graphs



Large and complex graphs capturing millions of entities and billions of relationships between entities.

Applications:

search, recommendation systems,
business intelligence, health
informatics, fact checking

Freebase : 1.9 billion triples

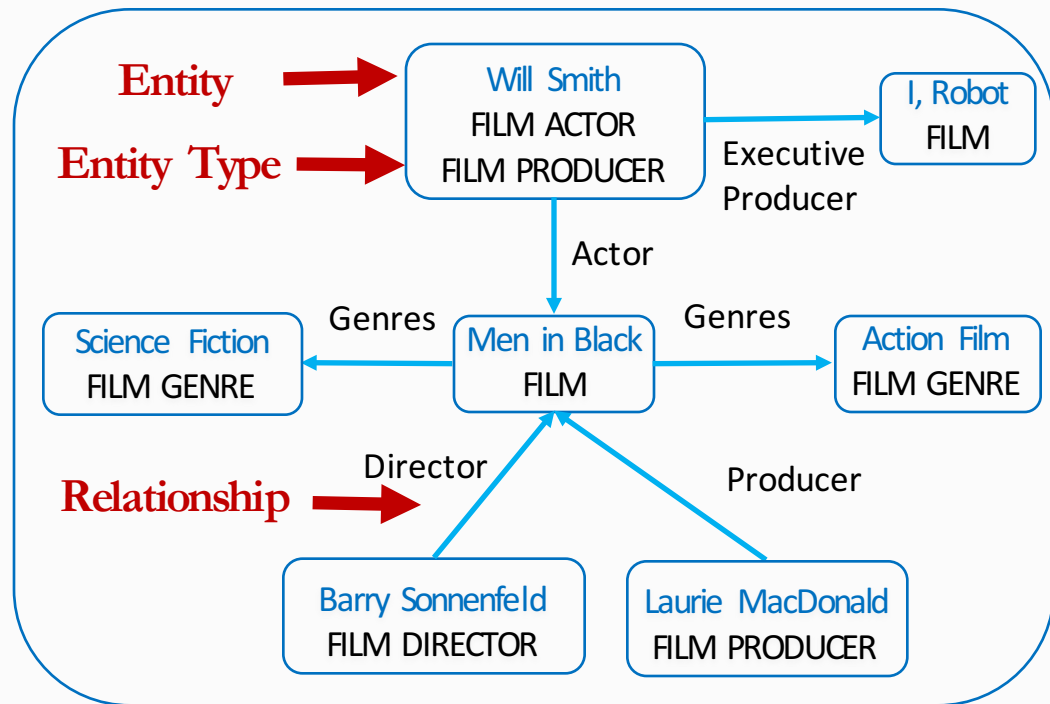
DBpedia : 3 billion triples

YAGO : 120 million triples

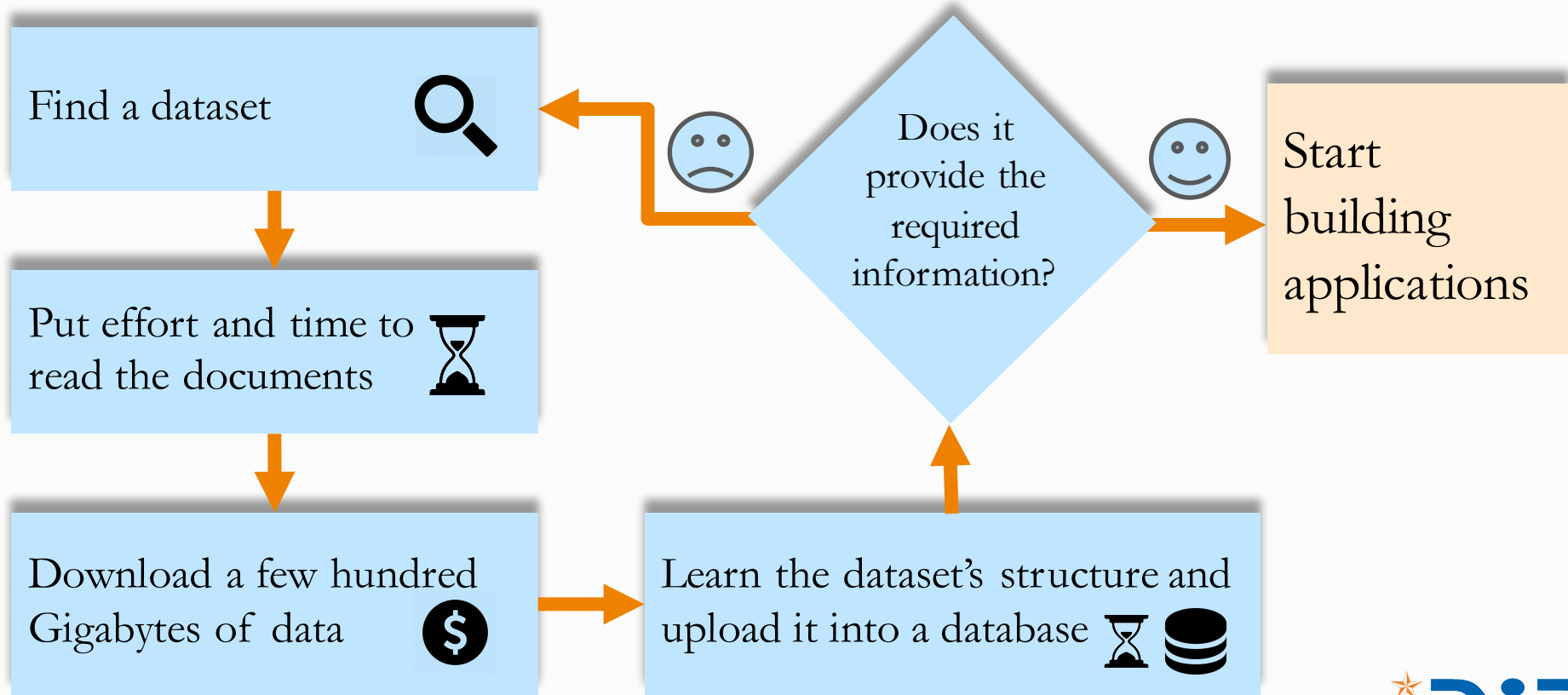
Linked Open Data :

hundreds of datasets

52 billion RDF triples



Steep Flag-Down Cost

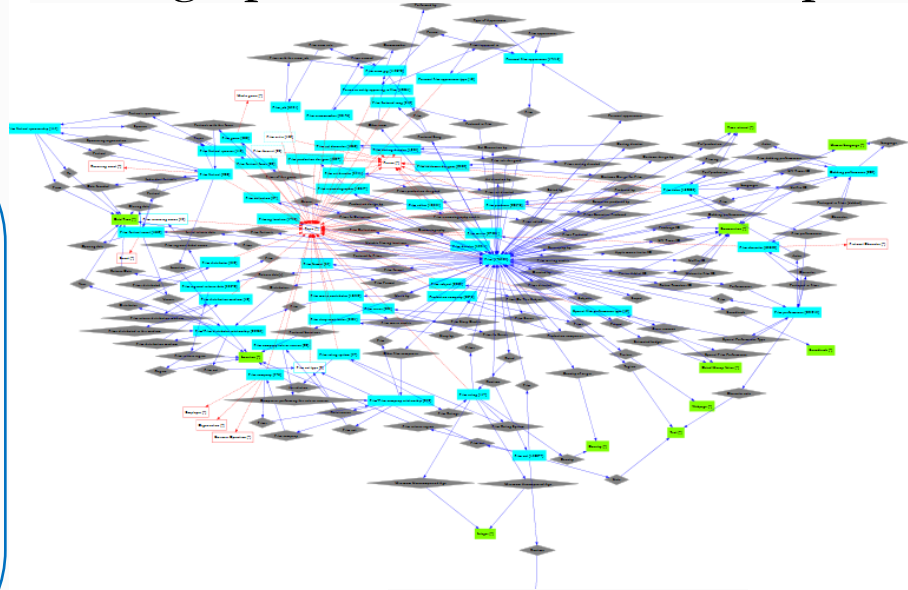
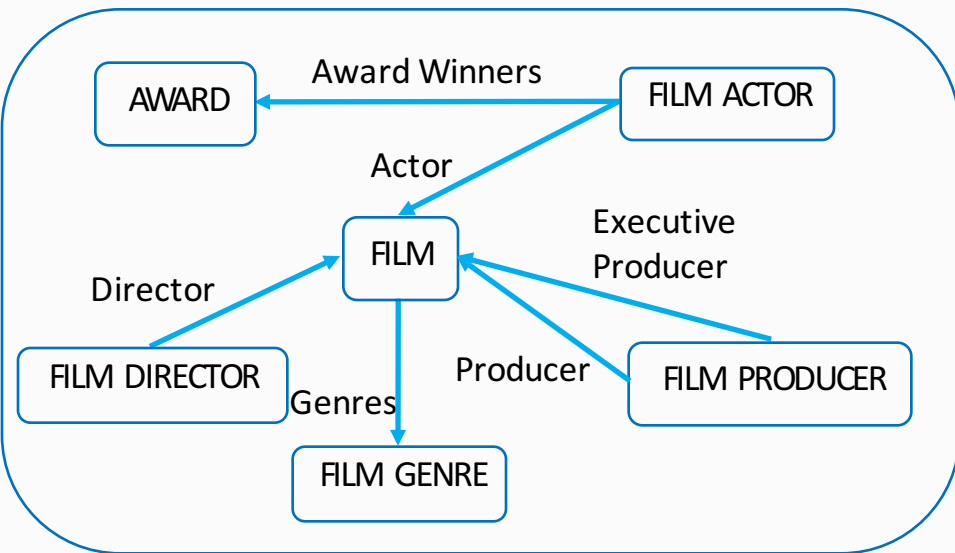


Need for a Quick Overview



Approach 1: Schema Graph

Schema graph itself can be too complex



Schema graph of “Film” domain in Freebase

Entity graph: 2M entities, 18 M edges

Schema graph: 63 entity types, 136 edges

Need for Quick Overview



Approach 2: Schema Summary

Schema summarization in relational database [Yang PVLDB09, Yang PVLDB11]

- Cluster tables in relational database by their semantic roles and similarities.
- Clusters tables, not relationships
- Detailed

XML summarization [Yu VLDB06]

- Provide a succinct overview of the entire schema graph

Graph summarization [Tian SIGMOD08, Zhang ICDE10]

- Group graph nodes based on their attribute similarity and allow users browse the summary from different grouping granularities.

Preview Tables

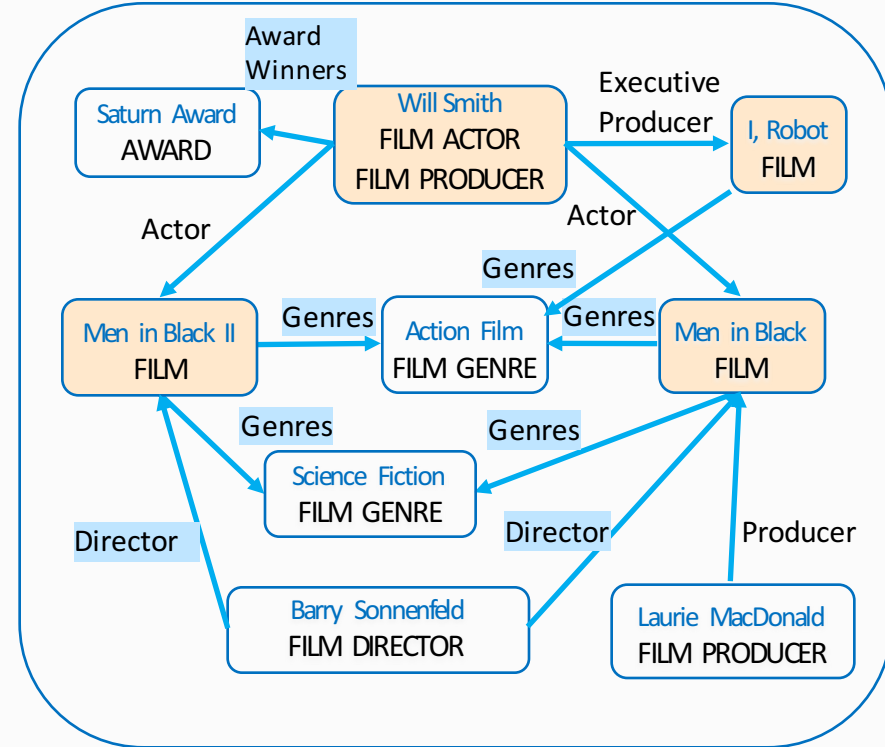


Key attributes

Non-key attributes

FILM ACTOR	Award Winners
Will Smith	Saturn Award

FILM	Director	Genres
Men in Black	Barry Sonnenfeld	{Action Film, Science Fiction}
Men in Black II	Barry Sonnenfeld	{Action Film, Science Fiction}
I, Robot	—	Action Film



Too Many Previews. Which one to Choose?

- Many possible previews
- Different choices

Preview A

FILM	FILM SET DECORATOR	FILM EDITOR
------	--------------------	-------------

PRODUCTION COMPANY	FILM
--------------------	------

Preview B

FILM	FILM DIRECTOR	FILM PRODUCER
------	---------------	---------------

FILM WRITER	FILM
-------------	------

Scoring Measures



FILM	Actor	Genres
4	6	5

FILM ACTOR	Actor	Award Winners
2	6	2

$$4 \times (6+5) = 44$$

$$2 \times (6+2) = 16$$

Score of the Preview

$$44 + 16 = 60$$

Key attribute scoring

- Coverage-based method
- Random walk-based method

Non-key attribute scoring

- Coverage-based method
- Entropy-based method

Optimal Preview Discovery



Find the preview with highest score that satisfies

- Size constraint
 - Number of key attributes K
 - Number of non-key attributes N
 - Distance between two preview tables d
- Concise**
- Tight** $\text{dist}(T_i, T_j) \leq d$
- Diverse** $\text{dist}(T_i, T_j) \geq d$

FILM	Performances	Genres	Directed By
------	--------------	--------	-------------

FILM DIRECTOR	Films Directed
---------------	----------------

FILM PRODUCER	Films Produced
---------------	----------------

Tight

FILM FESTIVAL	Location	Focus
---------------	----------	-------

FILM COMPANY	Films
--------------	-------

FILM CHARACTER	Portrayed in Film
----------------	-------------------

Diverse

Preview Discovery Algorithms



Concise preview

- Dynamic programming algorithm

Tight/Diverse preview

- NP-hard
- Apriori-like algorithm

Clique(G, k)



TightPreview ($G_s, k, k, 1, 0$)

DiversePreview ($G_s, k, k, 2, 0$)

Experiments



Dataset: Freebase

Accuracy of scoring measures

- Compared with Freebase ground truth:
 - Key attributes: Precision-at-k ($p@k$), Average Precision, Discounted Cumulative Gain (nDCG)
 - Non-key attributes: Mean Reciprocal Rank (MRR)
- Compared with crowd ranking:
 - Pearson Correlation Coefficient (PCC)

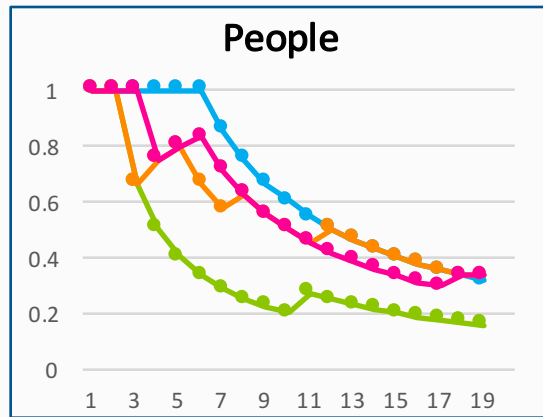
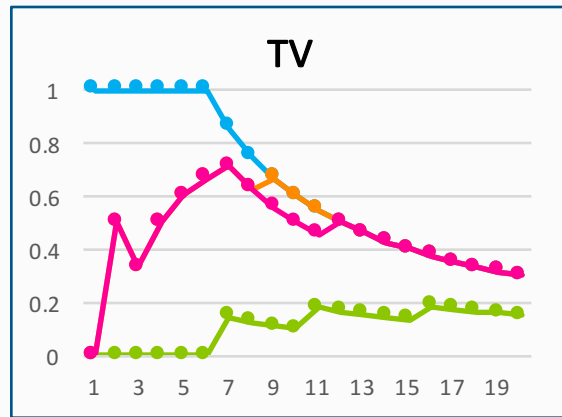
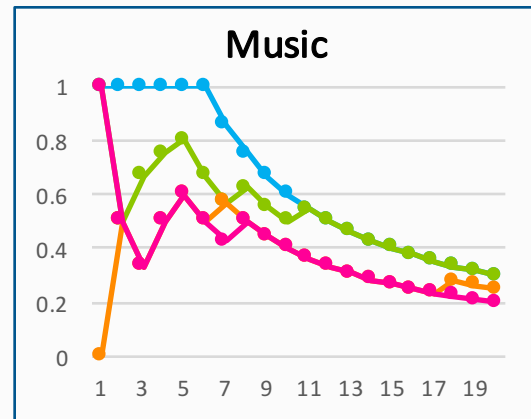
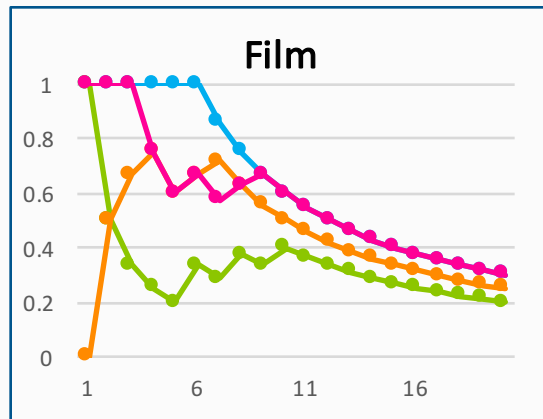
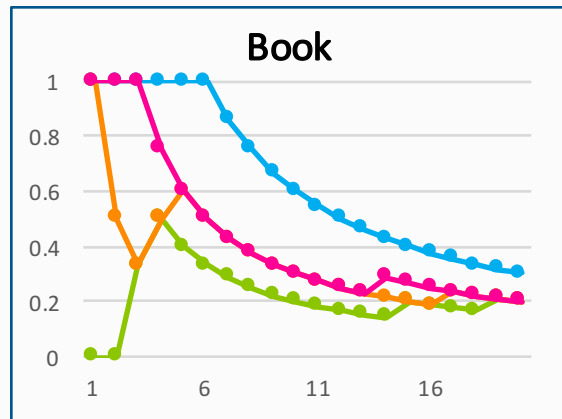
Efficiency of algorithms

- Execution time

User study

- Existence test questions
- User experience questions

Key Attribute Scoring ($p@k$)



- Optimal $p@k$
- YPS09 [Yang PVLDB09]
- Coverage
- Random Walk

User Study, Approaches Compared



- Domains:
film, books, music, TV, people
- Approaches:
 - Schema graph
 - Concise preview
 - Tight preview
 - Diverse preview
 - Freebase ground truth
 - YPS09
 - Hand-crafted preview tables
- 4 existence questions
- 4 experience questions
- 84 Master's and PhD students in database area
- \$15 gift card

1. Based on this schema summary, I know the dataset contains entities that are "film producer".

☐ Agree
 ☐ Not sure
 ☐ Disagree

Next

Film [View in a new Tab](#)

Film character	Portrayed in films
Iendil	actor: Peter McKenzie film: The Lord of the Rings: The Fellowship of the Ring
Foghorn Leghorn	actor: Mel Blanc film: False Hare special_performance, type: Voice
Strawberry Shortcake	actor: Russi Taylor film: Strawberry Shortcake Meets the Berrykins special_performance, type: Voice

Film actor	Film performances
Dusty Springfield	film: Music Scene: The Best of 1969-1970: Vol. 2
Cal Ripken, Jr.	film: Baseball: The Ripken Way: Pitching
Will Smith	character: Muhammad Ali film: Ali

Film director	Films directed
Jacobo Morales	Lo que le Paso a Santiago
Annie Leibovitz	Zoetrope
Starhawk	Full Circle

Film	Performances	Genres	Country of origin	Runtime
song of the Thin Man	actor: Myrna Loy character: Nora Charles	Thriller	United States of America	runtime:
The Hudsucker Proxy	actor: Tim Robbins character: Norville Barnes	Comedy	United States of America	runtime:
I Am Sam	actor: Sean Penn character: Sam Dawson	Courtroom Drama	United States of America	runtime:

Film crew member	Films crewed
Art Babbitt	None
Krishna	None
Shelby Young	film: My Soul to Take film_crew_role: Adr Voices

Film producer	Films Produced
Aamir Khan	Lagaan: Once Upon a Time in India
Ralph Fiennes	Coriolanus
Arthur Freed	An American in Paris

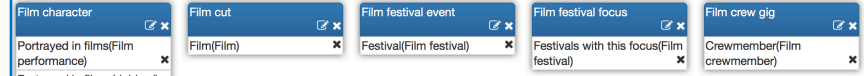
5. How easy was it to read the schema summary of this domain?

☐ 5. Very easy
 ☐ 4. Easy
 ☐ 3. Neutral
 ☐ 2. Hard
 ☐ 1. Very hard

Additional Comments

Next

- Individually and as a group



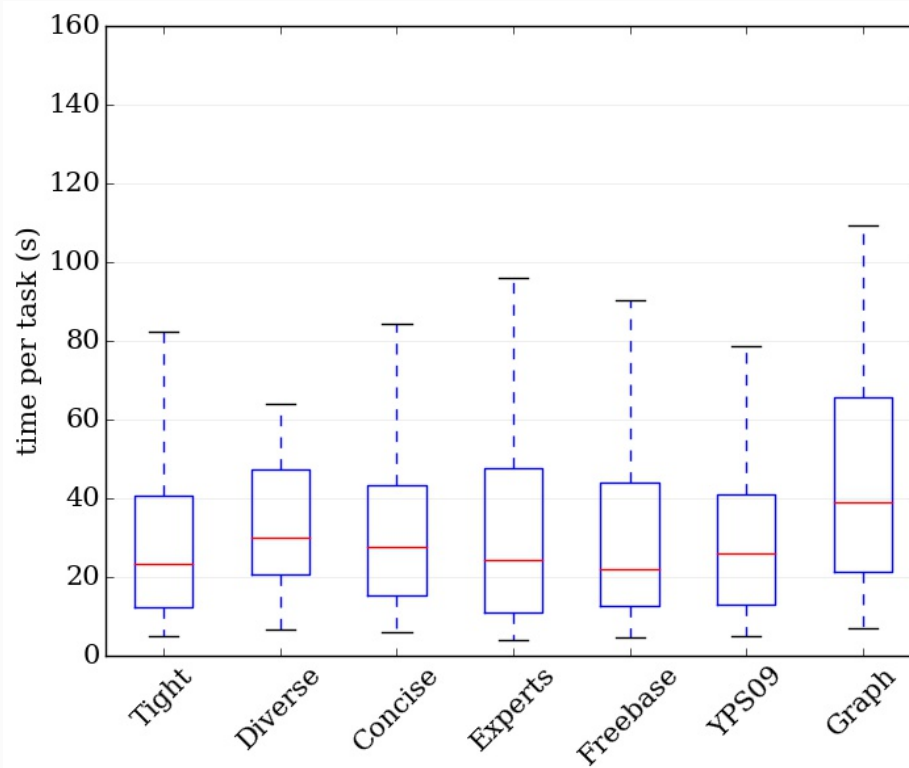
User Study: Existence Test Questions



	Tight	Diverse	Freebase	Experts	YPS09	Schema Graph
Concise	$z=1.59$ $p=0.0559$	$z=-2.28$ $p=0.0113$	$z=0.49$ $p=0.3121$	$z=-0.13$ $p=0.4483$	$z=0.36$ $p=0.3594$	$z=-0.43$ $p=0.3336$
Tight		$z=-3.48$ $p=0.0003$	$z=-1.12$ $p=0.1314$	$z=-1.69$ $p=0.0455$	$z=-1.282$ $p=0.0999$	$z=-1.93$ $p=0.0268$
Diverse			$z=2.57$ $p=0.0051$	$z=2.10$ $p=0.0179$	$z=2.60$ $p=0.0047$	$z=1.70$ $p=0.0446$
Freebase				$z=-0.61$ $p=0.2709$	$z=-0.15$ $p=0.4404$	$z=-0.87$ $p=0.1922$
Experts					$z=0.49$ $p=0.3121$	$z=-0.29$ $p=0.3859$
YPS09						$z=-0.77$ $p=0.2206$

Pairwise comparisons of conversion rates, domain="music", $\alpha=0.1$

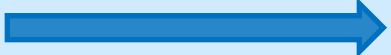
User Study: Existence Test Questions



Time taken on existence tests, domain="music"

User Study: User Experience Questions



Questions	most favorable  least favorable						
How easy was it to read the schema summary?	Freebase	Diverse	Graph	Experts	YPS09	Concise	Tight
How much understanding of the data can you gain from it?	Graph	Freebase	YPS09	Diverse	Concise	Tight	Experts
How helpful was it in assisting you to understand the data?	Graph	Freebase	YPS09	Diverse	Experts	Concise	Tight
Is it missing important information?	YPS09	Concise	Experts	Graph	Tight	Freebase	Diverse

Systems sorted by average user experience scores across five domains

Acknowledgment



UNIVERSITY OF
TEXAS
ARLINGTON



Disclaimer: This material is based upon work partially supported by the National Science Foundation Grants 1018865, 1408928 and the National Natural Science Foundation of China Grant 61370019. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the funding agencies.

Thank You! Questions?



Generating Preview Tables for Entity Graphs

Ning Yan, Sona Hasani, Abolfazl Asudeh, Chengkai Li