Generating Preview Tables for Entity Graphs

Ning Yan[1#], Sona Hasani[*], Abolfazl Asudeh[*], Chengkai Li[*]

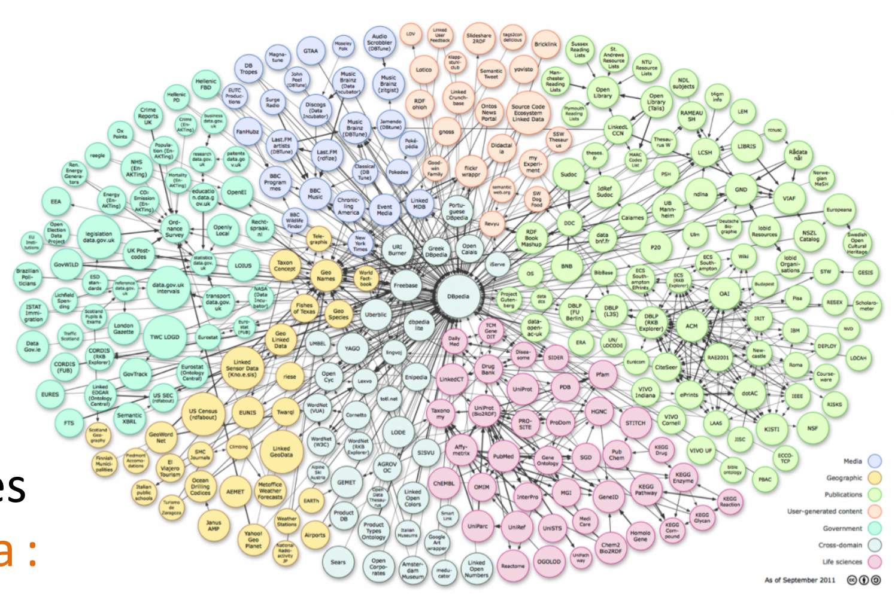
Huawei U.S. R&D Center[#] University of Texas at Arlington[*]

[1]The work was done while at UTA. Innovative Database and Information Systems Research (IDIR) Laboratory

Ultra-heterogeneous Entity Graphs

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

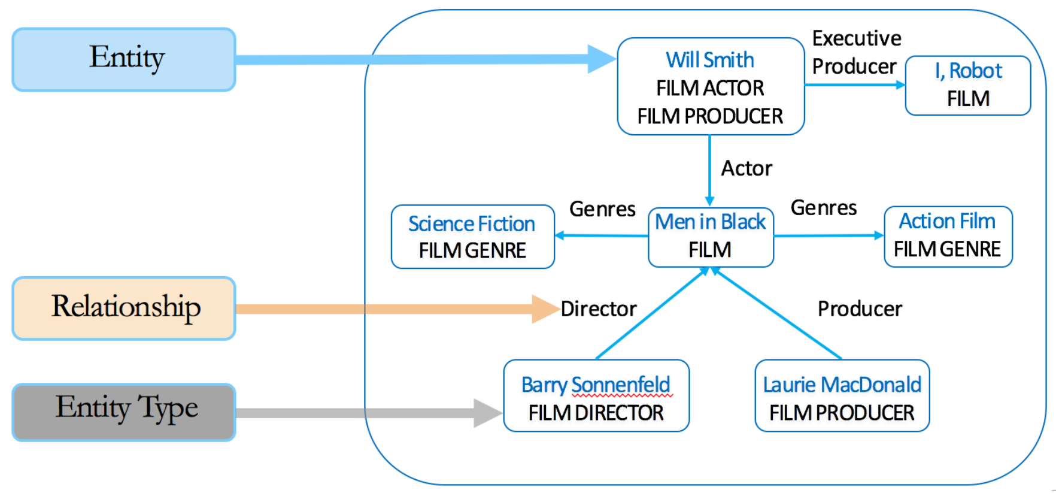
- ❑ Freebase : 1.9 billion triples
- ❑ DBpedia : 3 billion triples
- ❑ YAGO : 120 million triples
- ❑ Linked Open Data : 52 billion triples



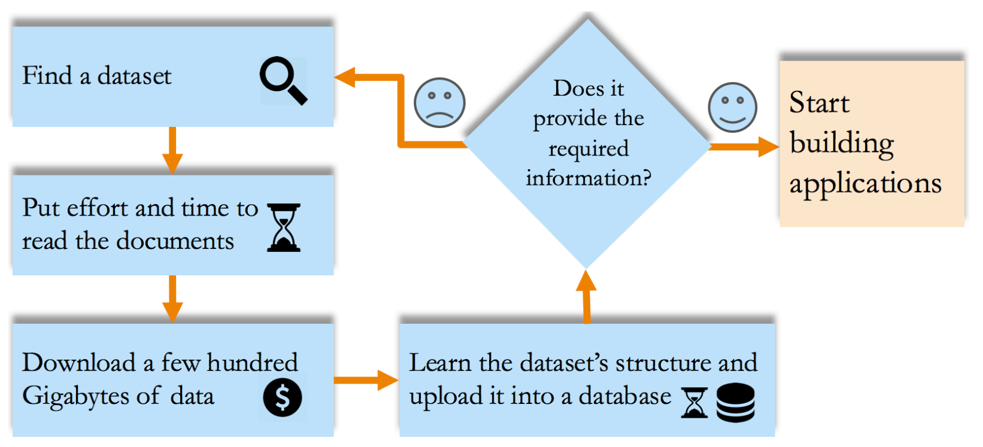
http://linkeddata.org/

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

Entity Graph



Steep Flag-Down Cost



Find a dataset

Put effort and time to read the documents

Download a few hundred Gigabytes of data

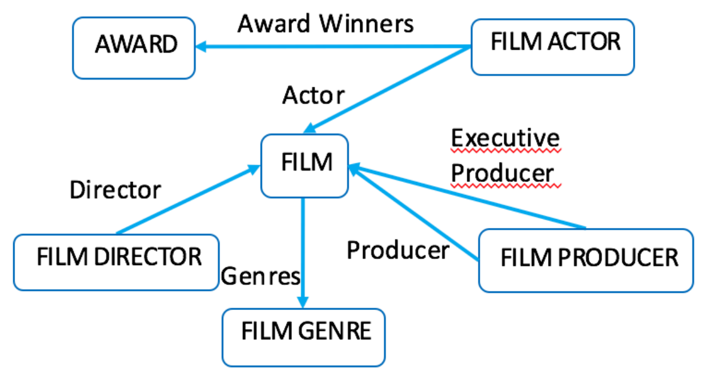
Learn the dataset's structure and upload it into a database

Does it provide the required information?

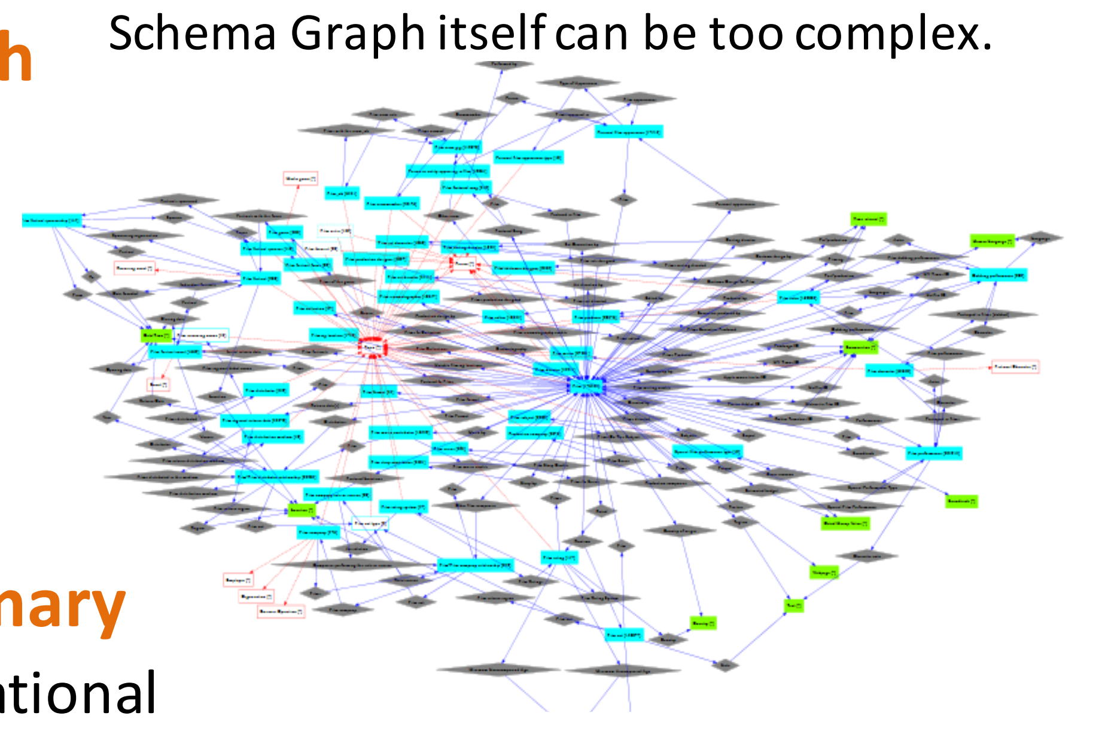
Start building applications

Need for a Quick Overview

Approach 1: Schema Graph



Schema Graph itself can be too complex.



Schema graph of "Film" domain in Freebase

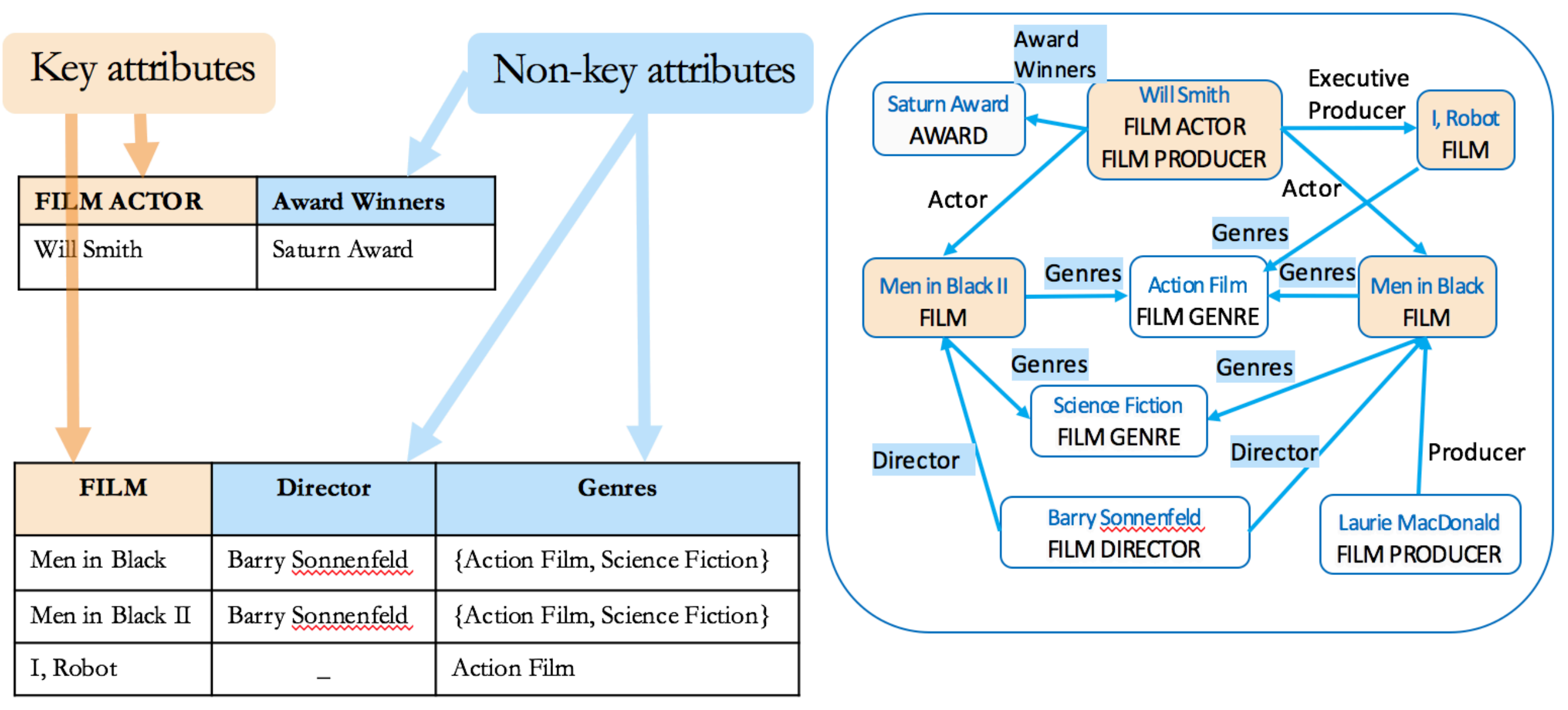
Entity graph:
2M entities, 18 M edges

Schema graph:
63 entity types ,136 edges

Approach 2: Schema Summary

- ❑ Schema summarization in relational database [Yang PVLDB09, Yang PVLDB11]
- ❑ XML summarization [Yu VLDB06]
- ❑ Graph summarization [Tian SIGMOD08, Zhang ICDE10]

Preview Tables



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?

A

FILM	FILM SET DECORATOR	FILM EDITOR

PRODUCTION COMPANY	FILM

B

FILM	FILM DIRECTOR	FILM PRODUCER

FILM WRITER	FILM

Aggregate Scoring

FILM	Actor	Genres
4	6	5

FILM ACTOR	Actor	Award Winners
2	6	2

Score of the Preview

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

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

$$+ \rightarrow 60$$

Attribute Scoring

Key attribute scoring

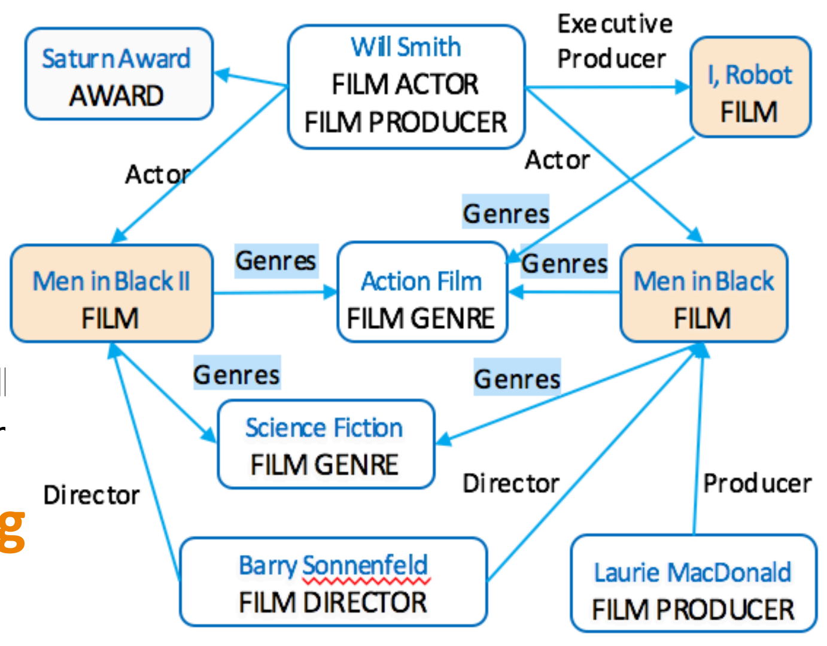
Coverage-based method:
Coverage(FILM) = 3

Random walk-based method :
Stationary distribution of a random walk process defined over the schema graph

Non-key attribute scoring

Coverage-based method :
Coverage(Genres) = 5

Entropy-based method :
Entropy(Genres) = (2/3) log(3/2)+(1/3) log(3/1) = 0.28



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}

Optimal Preview Discovery

Find the preview with highest score that satisfies

Concise
- Size constraint
  - Number of key attributes *K*
  - Number of non-key attributes *N*
- Distance between two preview tables *d*

Tight: dist(Ti, Tj) ≤ d

Diverse: dist(Ti, Tj) ≥ 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

Algorithms

Concise preview, dynamic programming algorithm

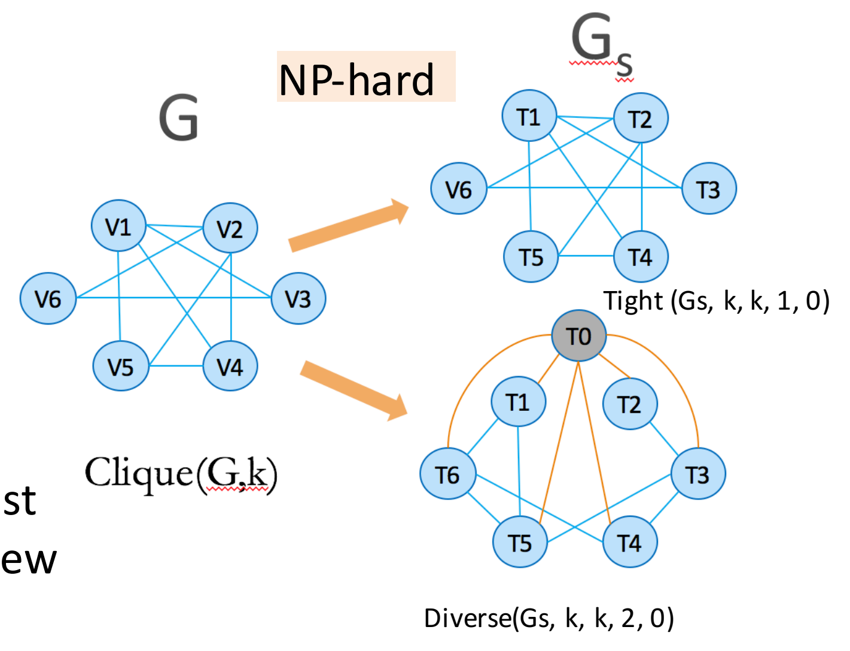
We assume all K key attributes are ordered arbitrarily.

optimal concise preview (k, n, X) is the best of:

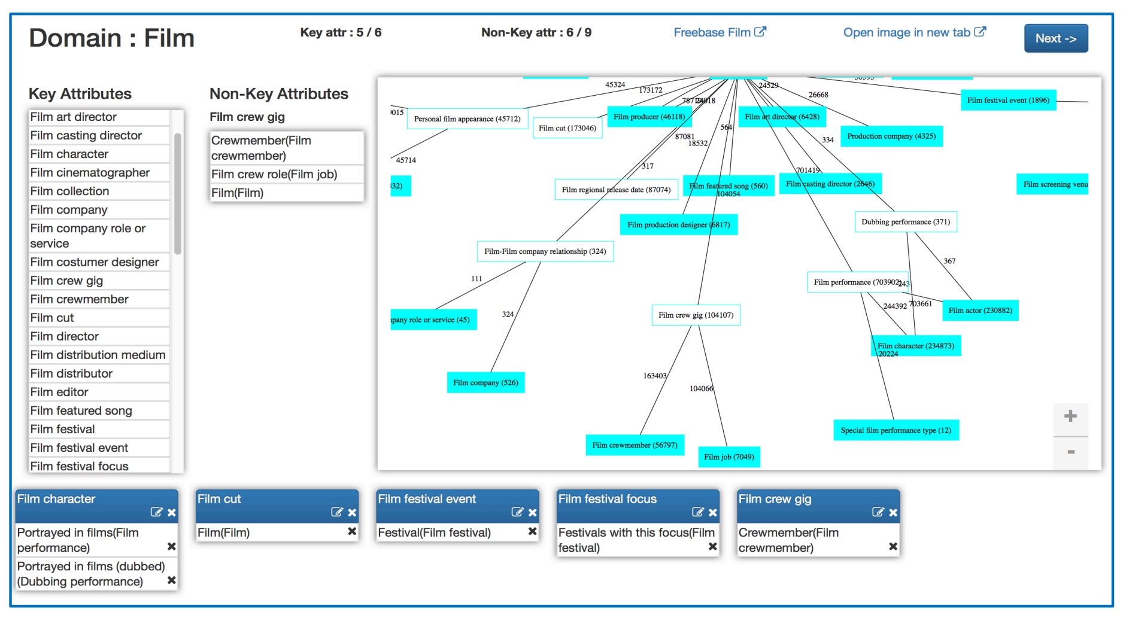
- optimal concise preview (k, n, X-1)
- optimal concise preview (k-1, n-1, X-1) ∪ X-th Key-attribute with 1 non-key attribute
- optimal concise preview (k-1, n-2, X-1) ∪ X-th Key-attribute with 2 non-key attributes
- … …
- optimal concise preview (k-1, k-1, X-1) ∪ X-th Key-attribute with (n-k+1) non-key attributes

Tight/Diverse preview, Apriori property algorithm

1. Construct 2-cliques by enumerating all key attribute pairs
2. for i = 3 to k
generate i-cliques from (i-1)-cliques based on Apriori property
3. find the k-clique with highest score, **return** as optimal preview



User Study



Domain : Film Key attr : 5 / 6 Non-Key attr : 6 / 9 Freebase Film Open image in new tab Next ->

Key Attributes: Film art director, Film casting director, Film character, Film cinematographer, Film collection, Film company, Film company role or service, Film costumer designer, Film crew gig, Film crewmember, Film cut, Film director, Film distribution medium, Film distributor, Film editor, Film featured song, Film festival, Film festival event, Film festival focus

Non-Key Attributes: Film crew gig: Crewmember(Film crewmember), Film crew role(Film job), Film(Film)

Film character: Portrayed in films(Film performance) ✖, Portrayed in films (dubbed) (Dubbing performance) ✖

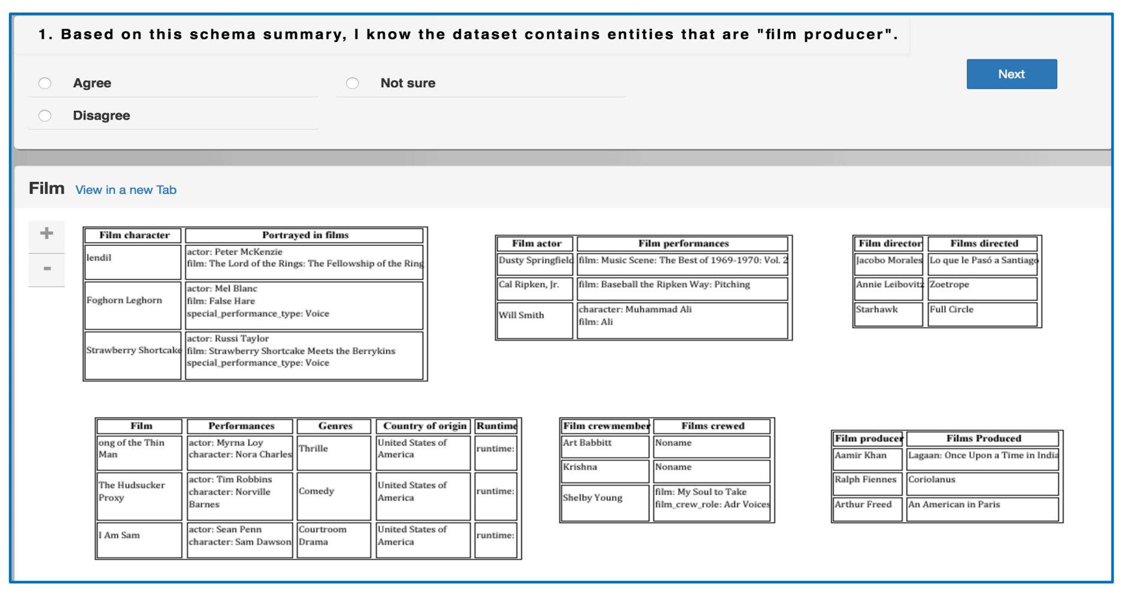
Film cut: Film(Film) ✖

Film festival event: Festival(Film festival) ✖

Film festival focus: Festivals with this focus(Film festival) ✖

Film crew gig: Crewmember(Film crewmember) ✖

Personal film appearance (45712)
Film cut (173046)
Film producer (46118)
Film art director (6428)
Production company (4325)
Film festival event (1896)
Film regional release date (87074)
Film featured song (560)
Film casting director (2646)
Film screening venue
Film production designer (6817)
Dubbing performance (371)
Film-Film company relationship (324)
Film performance (703902)
Film actor (230882)
Film crew gig (104107)
Film character (234873)
Film company (526)
Special film performance type (12)
Film crewmember (56797)
Film job (7049)
+
-



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
lendil	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 Pasó a Santiago
Annie Leibovitz	Zoetrope
Starhawk	Full Circle

Film	Performances	Genres	Country of origin	Runtime
ong 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 crewmember	Films crewed
Art Babbitt	Noname
Krishna	Noname
Shelby Young	film: My Soul to Take film_crew_role: Adr Voices

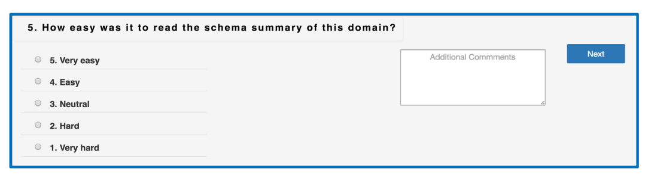
Film producer	Films Produced
Aamir Khan	Lagaan: Once Upon a Time in India
Ralph Fiennes	Coriolanus
Arthur Freed	An American in Paris

Domains:
film, books, music, TV, people

Hand-crafted preview tables
10 PhD students in Database research group
Individually and as a group
$20 gift card

Existence/experience questions
- ❑ Schema graph
- ❑ Concise preview
- ❑ Tight preview
- ❑ Diverse preview
- ❑ Freebase ground truth
- ❑ YPS09
- ❑ Hand-crafted preview tables

84 Master's and PhD students in database area
$15 gift card



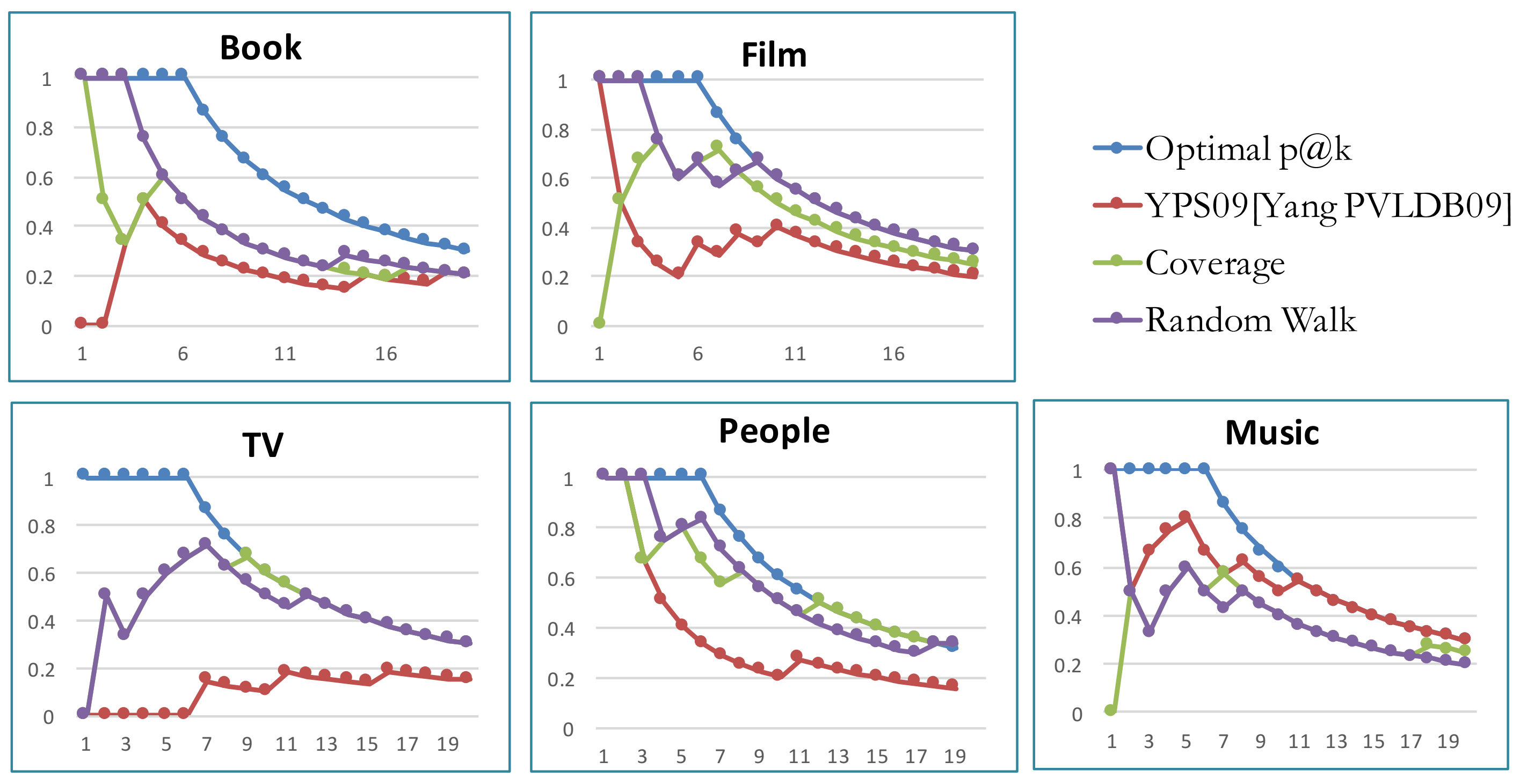
5. How easy was it to read the schema summary of this domain?

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

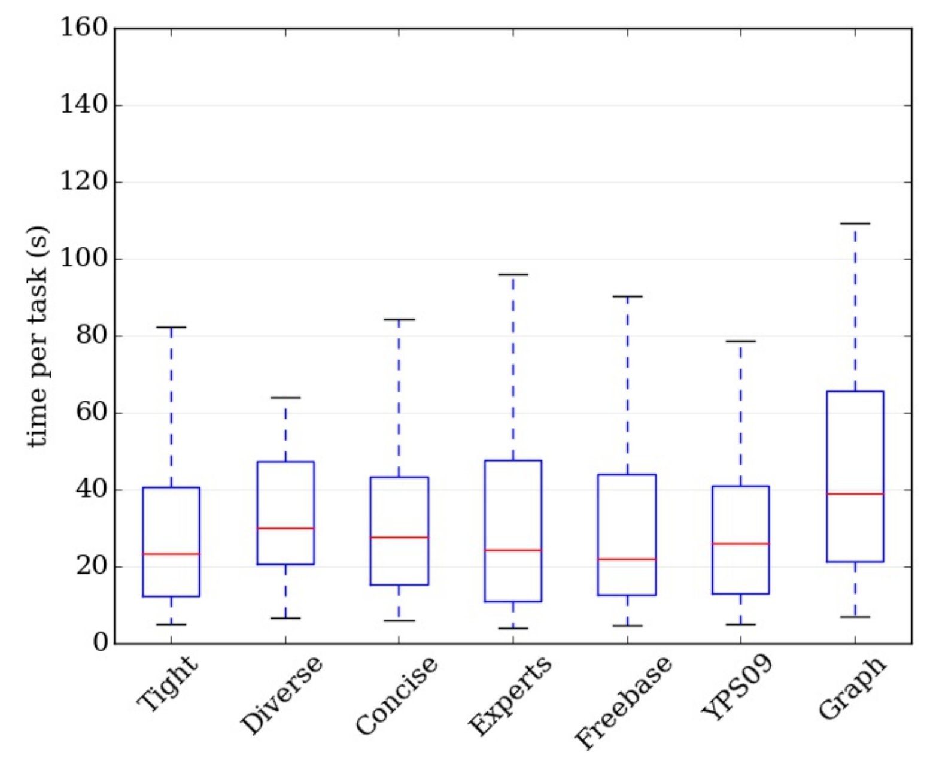
Additional Comments

Next

Experiment Results



Key attribute scoring (precision-at-k)



Time taken on existence tests, domain="music"

	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", α=0.1

	Key Attribute			Non-key Attribute	
Domain	YPS09	Coverage	Random Walk	Coverage	Entropy
books	0.4	0.55	0.43	0.43	0.43
film	-0.01	0.48	0.25	0.35	0.35
music	0.37	0.33	0.46	0.42	0.41
TV	0.37	0.69	0.65	0.47	0.47
people	0.36	0.31	0.29	0.43	0.43

Comparison between rankings by our approach and the crowd , Pearson Correlation Coefficient (PCC)

Domain	Coverage	Entropy
books	0.8	0.786
film	0.2	0.25
music	0.528	0.589
TV	0.622	0.379
people	0.708	0.606

Mean Reciprocal Rank (MRR) of Non-key attributes

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