Exploring Knowledge Bases through Questions and Faceted Answers

João Pedro V. Pinheiro[1], Marco A. Casanova[1], Elisa S. Menendez[2]

[1]Department of Informatics, PUC-Rio, Rio de Janeiro, RJ, Brazil

[2]Federal Institute of Education, Science and Technology of Bahia, Xique-Xique, BA, Brazil

{jpinheiro,casanova}@inf.puc-rio.br, elisa.menendez@ifbaiano.edu.br

http://www.inf.puc-rio.br/~jpinheiro/

Introduction

- Question Answering (QA) systems combine techniques from multiple fields of computer science, among which:
  - Natural Language Processing (NLP)
  - Information Retrieval
  - Machine Learning (ML)
  - Semantic Web

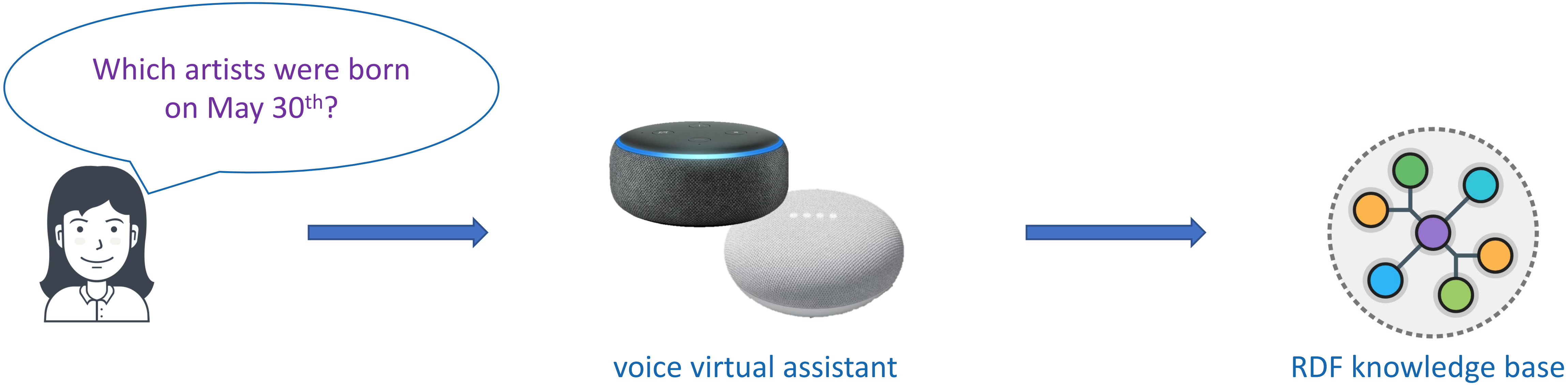
addressed in this dissertation

- A QA system may be split into two parts: *question*, which receives a user's input in natural language, transforms it into a structured query and searches the data; and *answer*, which displays consistent results in a human-readable format to the user

Introduction

- This study addresses the problem of *query answer modification* to improve the quality of the user's experience
  - in the context of an RDF knowledge base
- The dissertation proposes a fully automated process that reorganizes the original query answer by applying heuristics to summarize the results
- The heuristics together with a set of thresholds allow:
  - deciding which properties are interesting to apply aggregations (group by operations)
  - deciding if the answer is ready to be displayed to the user, or if the answer must be improved

Motivation

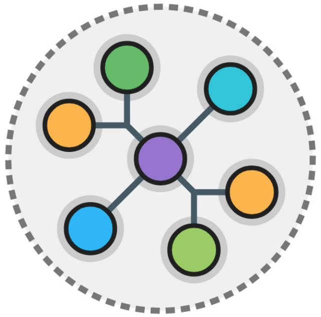


Which artists were born on May 30th?

voice virtual assistant

RDF knowledge base

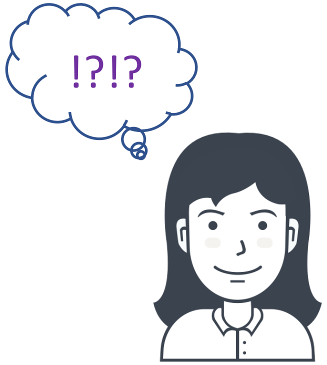
Motivation



RDF knowledge base

#	Artist	Background	Birth Date	Death Date	Gender	Nationality
1	Goodman, Benny	Jazz	1909-05-30	1986-06-13	Male	American
2	Leonhardt, Gustav	Classical	1928-05-30	2012-01-16	Male	Dutch
3	Green, CeeLo	Pop	1974-05-30	-	Male	American
4	Biosphere	Electronic	1962-05-30	-	Male	Norwegian
5	Fredriksson, Marie	Pop	1958-05-30	2019-12-09	Female	Swedish
…	…	…	…	…	…	…
122	Banhart, Devendra	Folk	1981-05-30	-	Male	American

very long result set

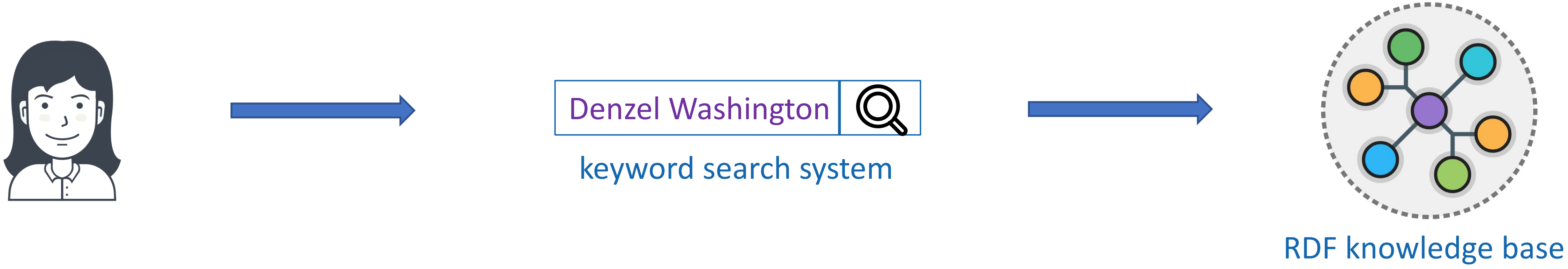


voice virtual assistant

Motivation

- Instead of listing the results, the virtual assistant may formulate questions to the user based on the prior result set, such as:
  - *“Do you want to list American or European artists?”*
  - *“Do you prefer Jazz, Pop, or Classical music?”*
  - *“Do you want to filter by active artists?”*

Motivation



Denzel Washington

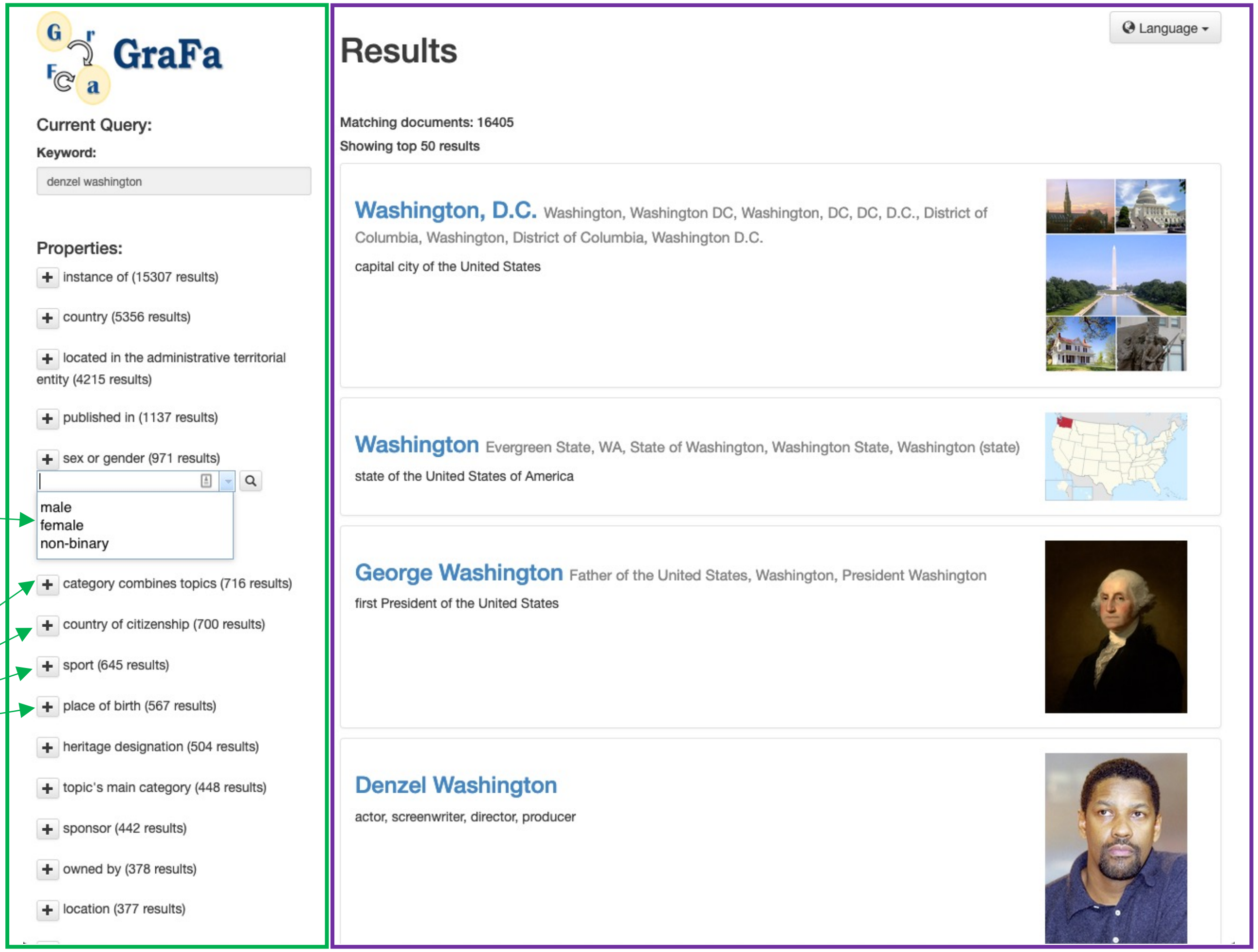
keyword search system

RDF knowledge base

Motivation

filter options

main results centered



GraFa

Current Query:

Keyword:

denzel washington

Properties:

- instance of (15307 results)
- country (5356 results)
- located in the administrative territorial entity (4215 results)
- published in (1137 results)
- sex or gender (971 results)
  - male
  - female
  - non-binary
- category combines topics (716 results)
- country of citizenship (700 results)
- sport (645 results)
- place of birth (567 results)
- heritage designation (504 results)
- topic's main category (448 results)
- sponsor (442 results)
- owned by (378 results)
- location (377 results)

Language

Results

Matching documents: 16405

Showing top 50 results

Washington, D.C. Washington, Washington DC, Washington, DC, DC, D.C., District of Columbia, Washington, District of Columbia, Washington D.C.

capital city of the United States

Washington Evergreen State, WA, State of Washington, Washington State, Washington (state)

state of the United States of America

George Washington Father of the United States, Washington, President Washington

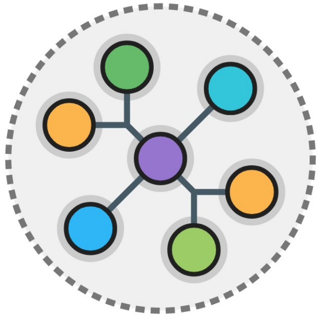
first President of the United States

Denzel Washington

actor, screenwriter, director, producer

facets

predicates



RDF knowledge base

Background and Related Work

Berners-Lee, T., Hendler, J., and Lassila, O. The semantic web; a new form of web content that is meaningful to computers will unleash a revolution of new possibilities. *Scientific American* (May 2001).

Linked Data (Semantic Web, RDF, SPARQL)

Webber, B. L. Questions, answers and responses: Interacting with knowledge-base systems. In *Topics in Information Systems*. Springer New York, 1986, pp. 365–402.

Question Answering (Humans, Machines)

Dalianis, H., and Hovy, E. Aggregation in natural language generation. In *Trends in Natural Language Generation An Artificial Intelligence Perspective*, J. G. Carbonell, J. Siekmann, G. Goos, J. Hartmanis, J. Leeuwen, G. Adorni, and M. Zock, Eds., vol. 1036. Springer Berlin Heidelberg, Berlin, Heidelberg, 1996, pp. 88–105.

Deutch, D., Frost, N., and Gilad, A. Provenance for natural language queries. *Proceedings of the VLDB Endowment 10*, 5 (Jan. 2017), 577–588.

Aggregation and Summarization

Moreno-Vega, J., and Hogan, A. Grafa: Faceted search & browsing for the wikidata knowledge graph. In *International Semantic Web Conference* (2018).

Faceted Browsing (or Faceted Search)

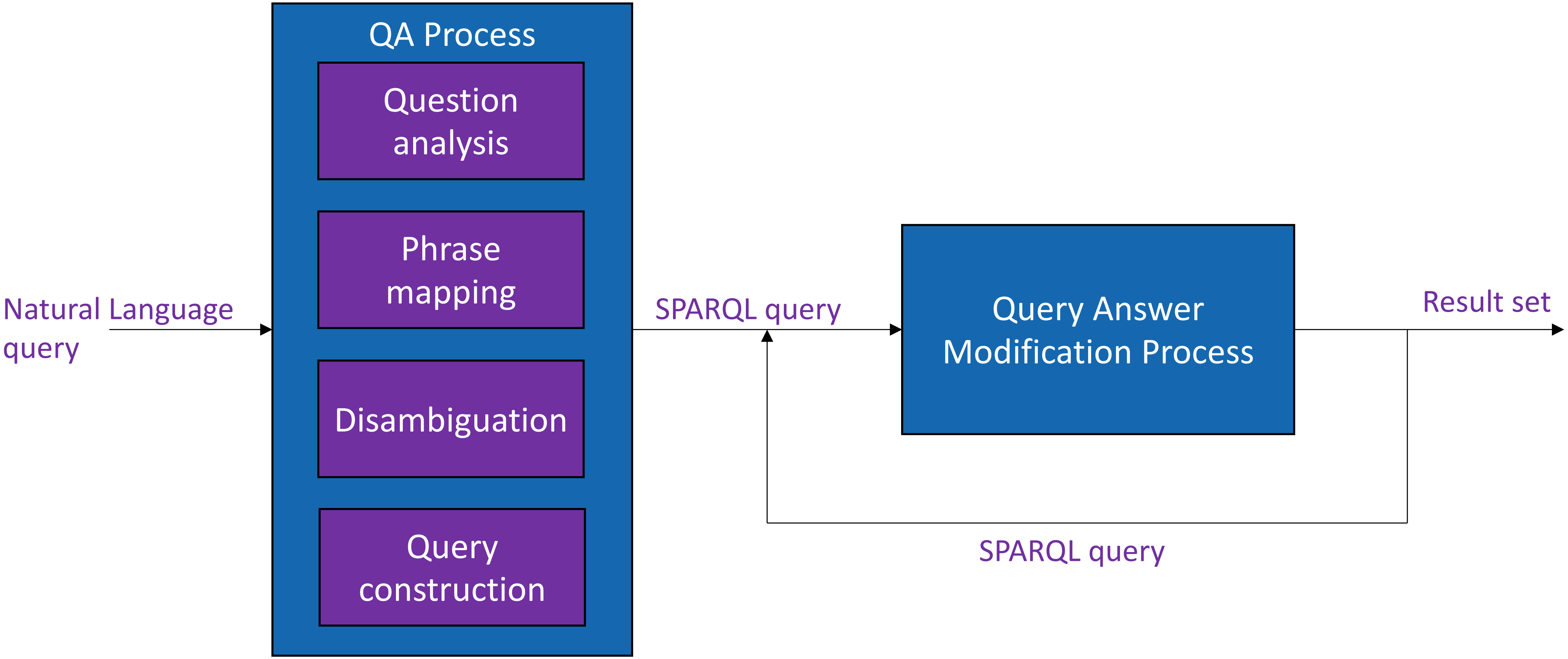
The Query Answer Modification Process

Overview

- Several studies addressed the problem of creating a question-answering (QA) interface to databases
- Usually, the proposed QA process has four steps:
  - Question Analysis
  - Phrase Mapping
  - Disambiguation
  - Query Construction
- The *query answer modification* process we propose starts after the query is executed

The Query Answer Modification Process

Overview



Natural Language query

QA Process

Question analysis

Phrase mapping

Disambiguation

Query construction

SPARQL query

Query Answer Modification Process

Result set

SPARQL query

The Query Answer Modification Process

Heuristics and Thresholds

- The process main goal is to reduce the size of the original result set using a set of heuristics with parameterized thresholds

- Heuristics responsibilities:
  - user's navigation thru data
  - automate the predicate/facet decision
- Thresholds responsibilities:
  - cut undesirable branches
  - reduce result set
  - control the stop condition

The Query Answer Modification Process

Heuristics and Thresholds

- The words **restrictive** and **embracing** are used to characterize predicates and facets

- In this context, their meaning are:
  - **restrictive**: reduce the total number of rows of the result set
  - **embracing**: keep or barely reduce the total number of rows of the result set

Initial Result Set: 49 rows

Predicate/Facet selected:
`imdb:label => Columbia/Tristar`

Final Result Set: 5 rows

Compression Rate: 89,80%

Initial Result Set: 91 rows

Predicate/Facet selected:
`imdb:color_info => Color`

Final Result Set: 74 rows

Compression Rate: 18,68%

The Query Answer Modification Process

Heuristics and Thresholds

- The heuristics tested were:
  - Σ : from the most **embracing** predicate, select the most **embracing** facet
  - ∏ : from the most **restrictive** predicate, select the most **embracing** facet
  - Ω : select the most **embracing** facet, regardless the predicate

- The thresholds applied were:
  - α : max number of predicates’ distinct values
  - β : max number of unique subjects to be returned to the user
  - δ : min and max range of predicates’ rate presence over unique subjects

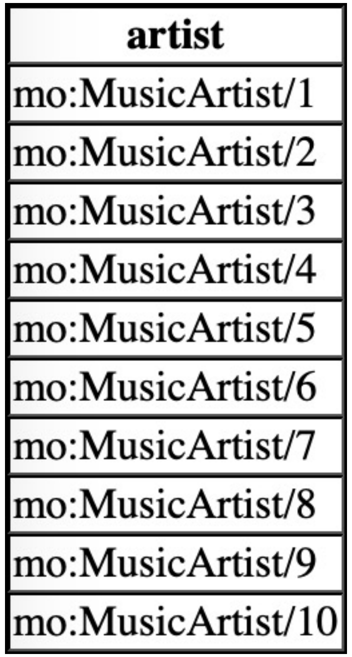
The Query Answer Modification Process

Transforming single-column into three-column result sets

- A simple approach for enriching the answers is to add to the instances returned their property values (denormalization)

```
prefix mo: <http://purl.org/ontology/mo/>
prefix foaf: <http://xmlns.com/foaf/0.1/>

select distinct ?artist
where {
        ?artist a mo:MusicArtist .
        ?artist dbo:birthDate ?date .
        filter(regex(?date, "5-30$", "i")) .
}
```



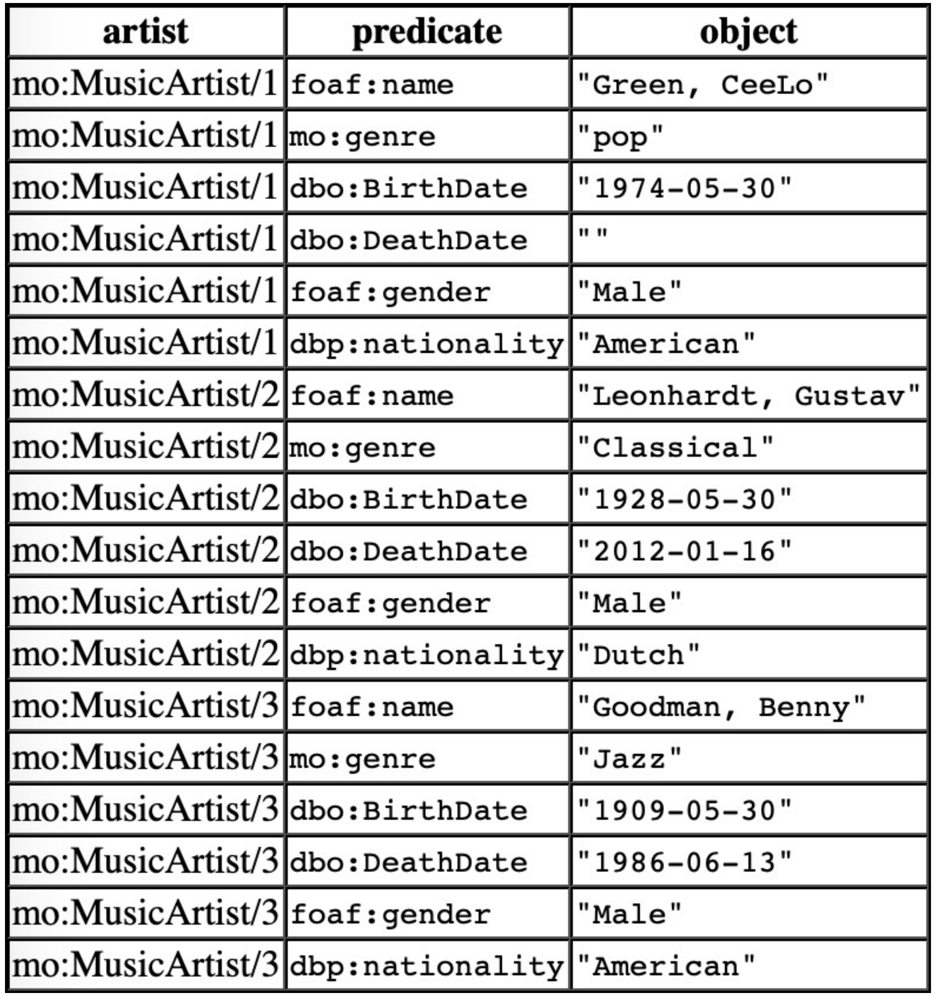
artist
mo:MusicArtist/1
mo:MusicArtist/2
mo:MusicArtist/3
mo:MusicArtist/4
mo:MusicArtist/5
mo:MusicArtist/6
mo:MusicArtist/7
mo:MusicArtist/8
mo:MusicArtist/9
mo:MusicArtist/10

The Query Answer Modification Process

Transforming single-column into three-column result sets

```
select distinct ?artist ?predicate ?object
where {
    {
        select ?artist
        where {
            ?artist a mo:MusicArtist .
            ?artist dbo:birthDate ?date .
            filter(regex(?date, "5-30$", "i")) .
        }

    } # graph pattern 1 – prior query
    . # conjunction (inner join)
    {
        select ?artist ?predicate ?object
        where {
            ?artist ?predicate ?object .
            filter(isLiteral(?object)) .
        }
    } # graph pattern 2
}
```



artist	predicate	object
mo:MusicArtist/1	foaf:name	"Green, CeeLo"
mo:MusicArtist/1	mo:genre	"pop"
mo:MusicArtist/1	dbo:BirthDate	"1974-05-30"
mo:MusicArtist/1	dbo:DeathDate	""
mo:MusicArtist/1	foaf:gender	"Male"
mo:MusicArtist/1	dbp:nationality	"American"
mo:MusicArtist/2	foaf:name	"Leonhardt, Gustav"
mo:MusicArtist/2	mo:genre	"Classical"
mo:MusicArtist/2	dbo:BirthDate	"1928-05-30"
mo:MusicArtist/2	dbo:DeathDate	"2012-01-16"
mo:MusicArtist/2	foaf:gender	"Male"
mo:MusicArtist/2	dbp:nationality	"Dutch"
mo:MusicArtist/3	foaf:name	"Goodman, Benny"
mo:MusicArtist/3	mo:genre	"Jazz"
mo:MusicArtist/3	dbo:BirthDate	"1909-05-30"
mo:MusicArtist/3	dbo:DeathDate	"1986-06-13"
mo:MusicArtist/3	foaf:gender	"Male"
mo:MusicArtist/3	dbp:nationality	"American"

The Query Answer Modification Process

Frequency analysis based on computed metadata

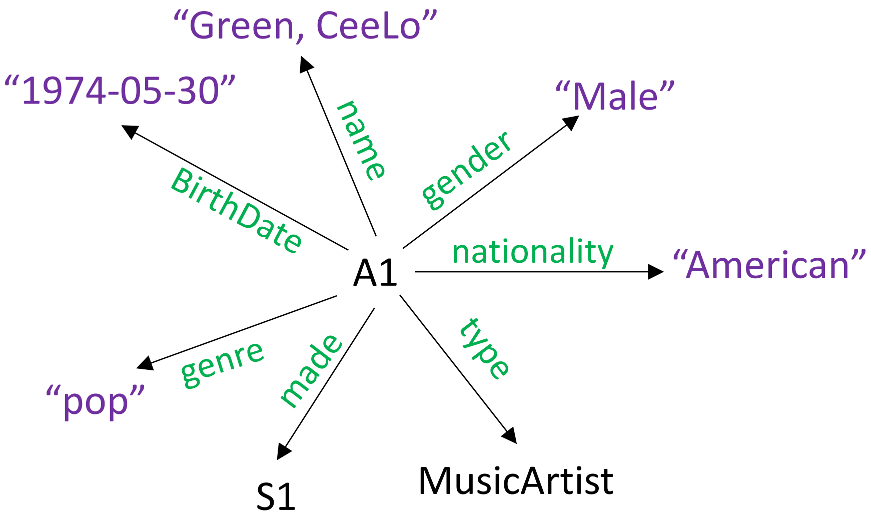
- There are two types of frequencies used in the process
  - *global frequency*: defined over full graph and computed once
  - *local frequency*: defined over sub-graph and computed at run time
  - both *frequencies* are computed over predicates pointing to literals only
- Entity ranking is based on *InfoRank*, a family of importance measures proposed in *Menendez et al. (2019)*
  - approach inspired by *PageRank* algorithm to propagate the importance scores from entity to entity
  - helps our process prioritize the most relevant triples of the result set

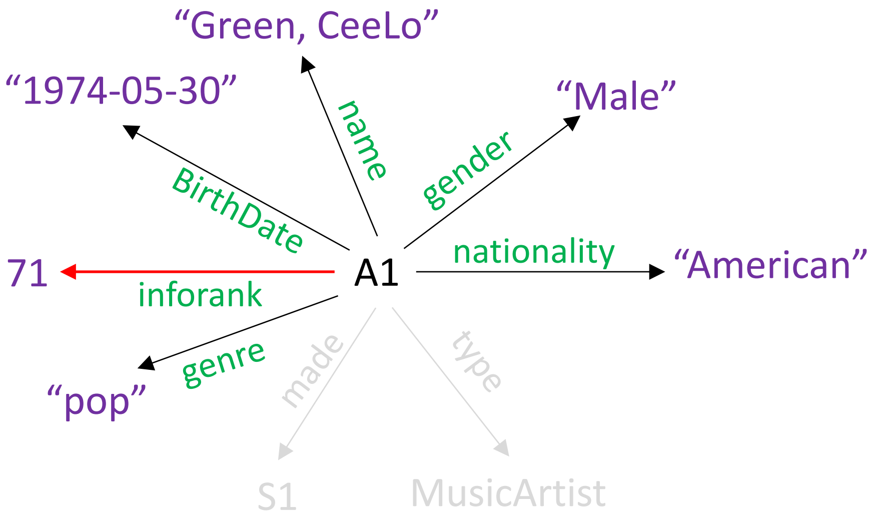
Menendez, E. S., Casanova, M. A., Leme, L. A. P., and Boughanem, M. (2019). Novel node importance measures to improve keyword search over rdf graphs. In International Conference on Database and Expert Systems Applications, pages 143–158. Springer. DOI:https://doi.org/10.1007/978-3-030-27618-8_11

The Query Answer Modification Process

Frequency analysis based on computed metadata

- Example with instance **A1**
  - left figure is the initial state: pointing to literals and other instances
  - right figure is the final state: pointing to literals only and *inforank* score
- The parameterized threshold δ is used to filter predicates that are candidates to be used in a group by operation





The Query Answer Modification Process

Verifying stop condition

- The process stops when:
  - the number of unique subjects is **less or equal than** the threshold β; **or**
  - the previous result set is **the same as** the current result set; **and**
  - there is an empty set of selectable predicates
- Also, the final result set is sorted by *InfoRank*

Experiments

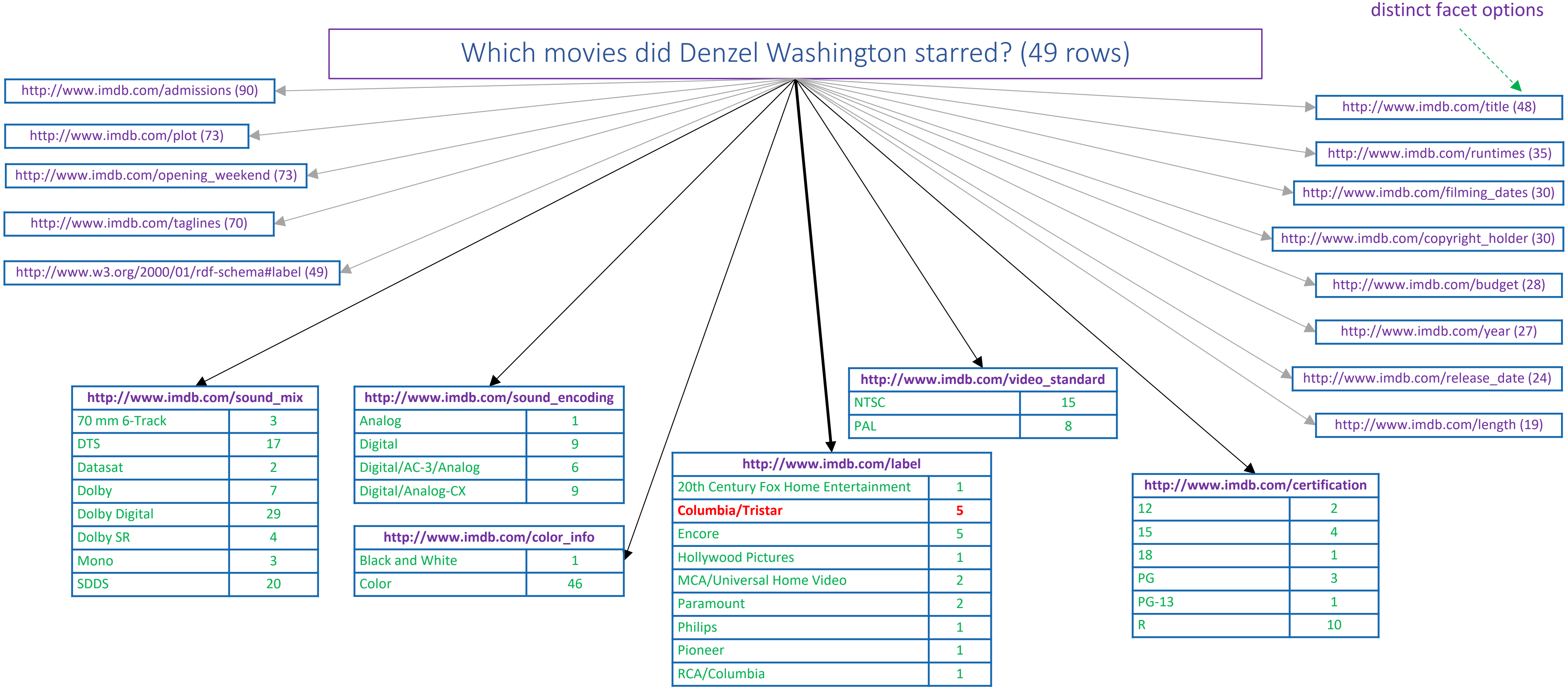
Benchmarks

- We performed initial experiments using the RDF versions of MusicBrainz and IMDb datasets (~200MM tuples each)
    - MusicBrainz dump enriched with DBpedia data
    - IMDb relational data transformed into RDF through R2RML
- We used sample queries from the QALD challenge and Coffman's benchmark
- To simplify the visualization, we decided to illustrate the example flows as K-D trees
    - the examples were generated applying heuristic Σ

Experiments

K-D Three Example with IMDb dataset – 1st step

Σ : most **embracing** predicate, most **embracing** facet
α = 10 : max number of predicates’ distinct values (distinct facet options)
β = 15 : max number of unique subjects in final result set
δ = (0.4, 2) : min and max range of predicates’ rate presence over unique subjects



distinct facet options

Which movies did Denzel Washington starred? (49 rows)

http://www.imdb.com/admissions (90)

http://www.imdb.com/plot (73)

http://www.imdb.com/opening_weekend (73)

http://www.imdb.com/taglines (70)

http://www.w3.org/2000/01/rdf-schema#label (49)

http://www.imdb.com/title (48)

http://www.imdb.com/runtimes (35)

http://www.imdb.com/filming_dates (30)

http://www.imdb.com/copyright_holder (30)

http://www.imdb.com/budget (28)

http://www.imdb.com/year (27)

http://www.imdb.com/release_date (24)

http://www.imdb.com/length (19)

http://www.imdb.com/sound_mix	
70 mm 6-Track	3
DTS	17
Datasat	2
Dolby	7
Dolby Digital	29
Dolby SR	4
Mono	3
SDDS	20

http://www.imdb.com/sound_encoding	
Analog	1
Digital	9
Digital/AC-3/Analog	6
Digital/Analog-CX	9

http://www.imdb.com/color_info	
Black and White	1
Color	46

http://www.imdb.com/label	
20th Century Fox Home Entertainment	1
Columbia/Tristar	5
Encore	5
Hollywood Pictures	1
MCA/Universal Home Video	2
Paramount	2
Philips	1
Pioneer	1
RCA/Columbia	1

http://www.imdb.com/video_standard	
NTSC	15
PAL	8

http://www.imdb.com/certification	
12	2
15	4
18	1
PG	3
PG-13	1
R	10

Σ : most **embracing** predicate, most **embracing** facet

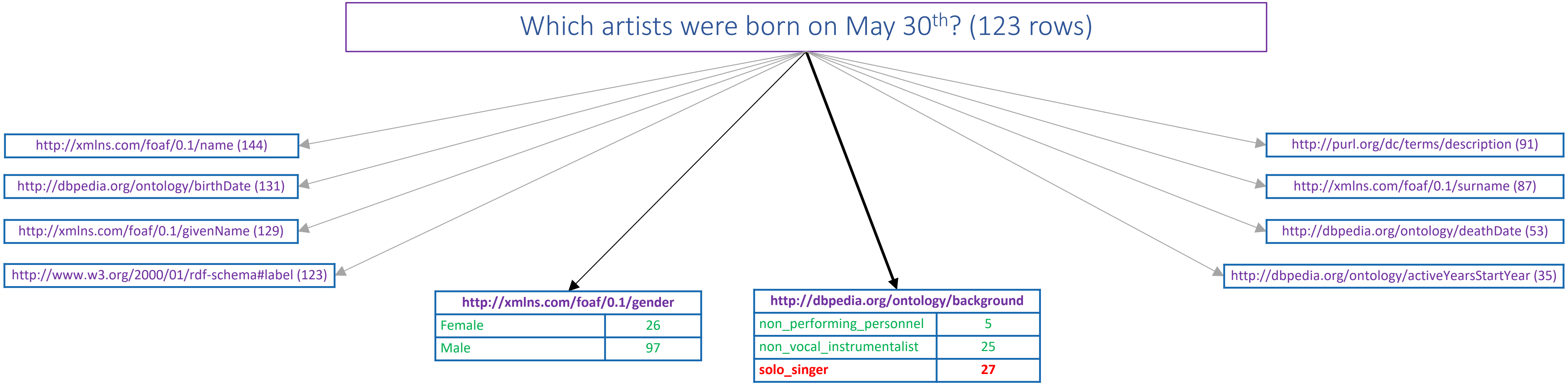
α = 10 : max number of predicates’ distinct values (distinct facet options)

β = 15 : max number of unique subjects in final result set

δ = (0.4, 2) : min and max range of predicates’ rate presence over unique subjects

Experiments

K-D Three Example with MusicBrainz dataset – 1st step



Which artists were born on May 30th? (123 rows)

http://xmlns.com/foaf/0.1/name (144)

http://dbpedia.org/ontology/birthDate (131)

http://xmlns.com/foaf/0.1/givenName (129)

http://www.w3.org/2000/01/rdf-schema#label (123)

http://purl.org/dc/terms/description (91)

http://xmlns.com/foaf/0.1/surname (87)

http://dbpedia.org/ontology/deathDate (53)

http://dbpedia.org/ontology/activeYearsStartYear (35)

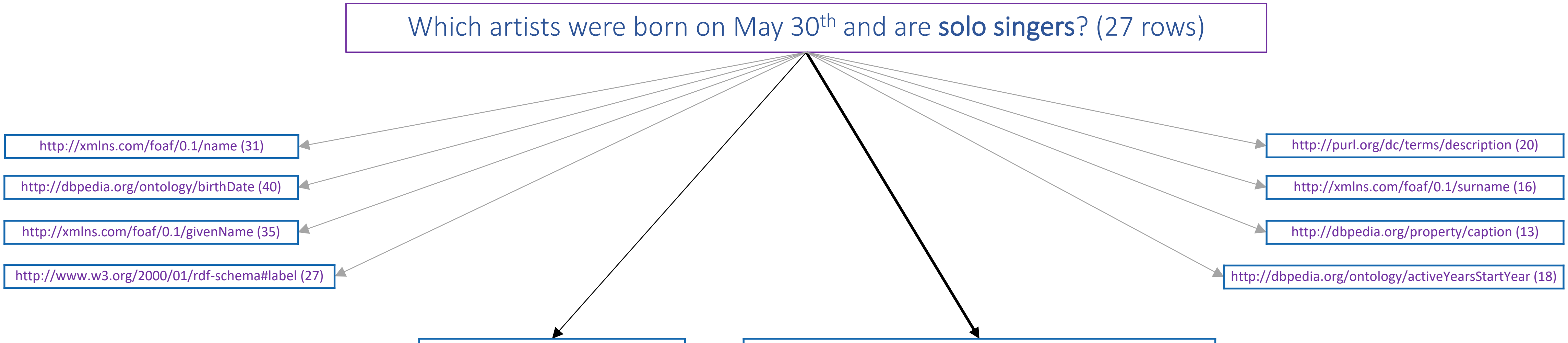
http://xmlns.com/foaf/0.1/gender	
Female	26
Male	97

http://dbpedia.org/ontology/background	
non_performing_personnel	5
non_vocal_instrumentalist	25
solo_singer	**27**

Σ : most **embracing** predicate, most **embracing** facet
α = 10 : max number of predicates’ distinct values (distinct facet options)
β = 15 : max number of unique subjects in final result set
δ = (0.4, 2) : min and max range of predicates’ rate presence over unique subjects

Experiments

K-D Three Example with MusicBrainz dataset – 2nd step



Which artists were born on May 30th and are **solo singers**? (27 rows)

http://xmlns.com/foaf/0.1/name (31)
http://dbpedia.org/ontology/birthDate (40)
http://xmlns.com/foaf/0.1/givenName (35)
http://www.w3.org/2000/01/rdf-schema#label (27)
http://purl.org/dc/terms/description (20)
http://xmlns.com/foaf/0.1/surname (16)
http://dbpedia.org/property/caption (13)
http://dbpedia.org/ontology/activeYearsStartYear (18)

http://xmlns.com/foaf/0.1/gender	
Female	7
Male	20

http://dbpedia.org/property/occupation	
Musician	1
Musician, singer-songwriter, record label owner	1
Musician, songwriter	1
Singer	2
Singer, actor	1
Singer, author, philanthropist, actress	1
Singer, rapper, songwriter, record producer, actor, businessman	1
Singer-songwriter	1
Singer-songwriter, musician	**5**
Singer-songwriter, musician, visual artist	1

Experiments

Compression Rate and final analysis

- A *Compression Rate* metric was defined and used to compare the obtained results

1. γ = compression rate
2. η = # lines of the initial result set
3. κ = # lines of the final result set
4. $\gamma = 1 - \kappa / \eta$

when $\kappa \sim \eta$ the compression is low

when $\kappa <<< \eta$ the compression is high

- Recall that variable β sets an upper bound on the final result set number of lines

Experiments

Compression Rate and final analysis - MusicBrainz

- Ten questions were originally tested over dataset, but three of them had not a single selectable predicate available

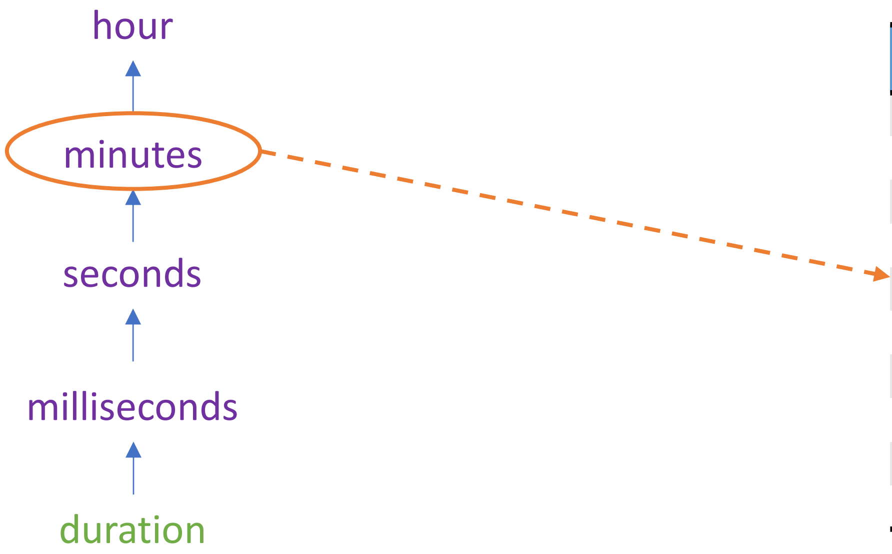
- Question: *“What are the songs performed by Aretha Franklin?”*
- Predicates:
  - `mo:track_number` (54 distinct facet options)
  - `rdfs:label` (1120 distinct facet options)
  - `mo:duration` (1881 distinct facet options)

- All predicates had the number of facets greater than α

Experiments

Compression Rate and final analysis - MusicBrainz

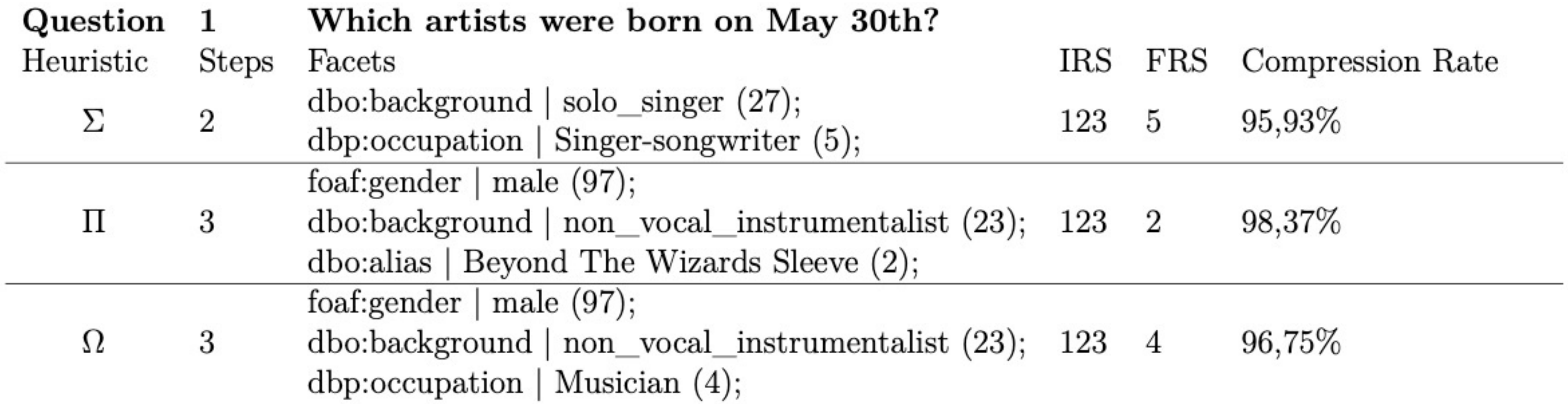
- Suppose the process have taxonomy information about specific predicates
  - the aggregations would be applied over multi-level
  - `mo:duration` is represented in milliseconds



#	Duration	Total Songs
1	11.0	9
2	10.0	6
3	9.0	21
4	8.0	28
5	7.0	69
6	6.0	157
7	5.0	337
8	4.0	781
9	3.0	1110
10	2.0	399

Experiments

Compression Rate and final analysis - MusicBrainz



Question	**1**	**Which artists were born on May 30th?**			
Heuristic	Steps	Facets	IRS	FRS	Compression Rate

Σ	2	dbo:background | solo_singer (27);
		dbp:occupation | Singer-songwriter (5);
			123	5	95,93%

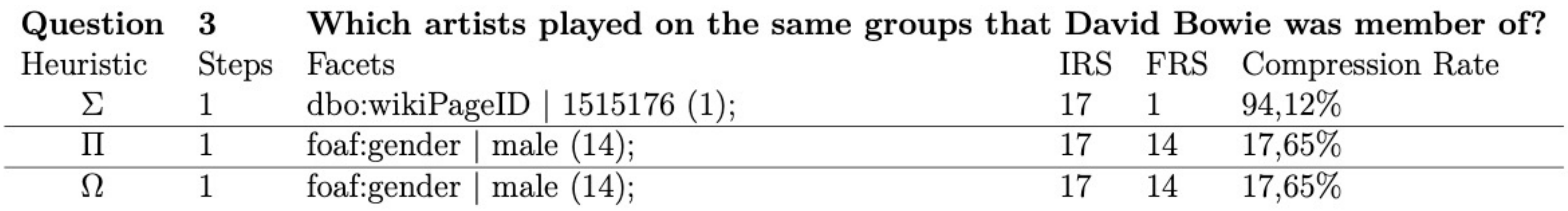
Π	3	foaf:gender | male (97);
		dbo:background | non_vocal_instrumentalist (23);
		dbo:alias | Beyond The Wizards Sleeve (2);
			123	2	98,37%

Ω	3	foaf:gender | male (97);
		dbo:background | non_vocal_instrumentalist (23);
		dbp:occupation | Musician (4);
			123	4	96,75%

- The first question presented excellent *Compression Rate* and selected meaningful predicates/facets
- Only the Π heuristic chose a questionable predicate in its 3rd step
  - `dbo:alias` should be almost unique by artist
  - the process is selecting a specific artist instead of selecting a property that few artists have in common

Experiments

Compression Rate and final analysis - MusicBrainz

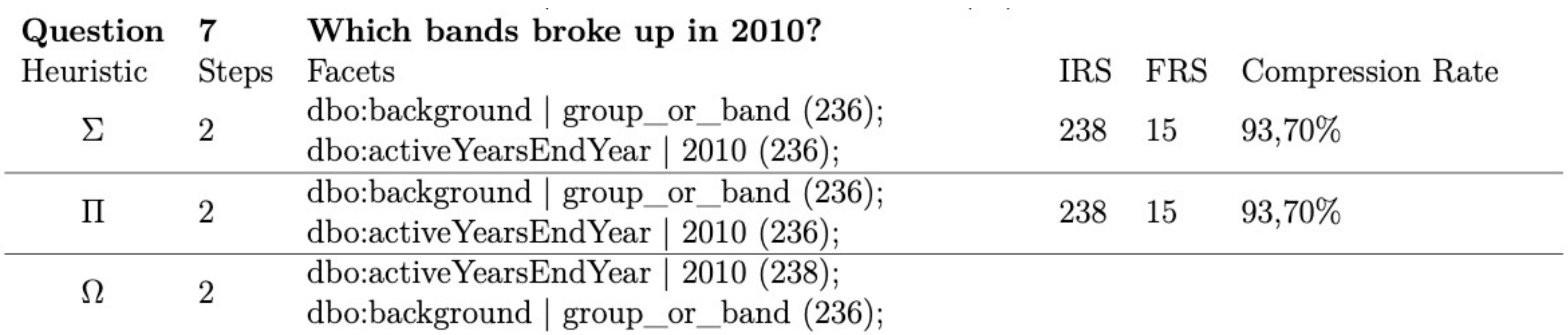


Question	**3**	**Which artists played on the same groups that David Bowie was member of?**			
Heuristic	Steps	Facets	IRS	FRS	Compression Rate
Σ	1	dbo:wikiPageID | 1515176 (1);	17	1	94,12%
Π	1	foaf:gender | male (14);	17	14	17,65%
Ω	1	foaf:gender | male (14);	17	14	17,65%

- In this case, the initial result set was already very close to variable β
- The *Compression Rate* from heuristics Π and Ω were the worst
- Analyzing the predicate’s meaning, heuristic Σ made a bad choice
  - `foaf:gender` seems much more interesting than `dbo:wikiPageID`
  - the *Compression Rate* metric by itself does not mean all

Experiments

Compression Rate and final analysis - MusicBrainz



Question	**7**	**Which bands broke up in 2010?**			
Heuristic	Steps	Facets	IRS	FRS	Compression Rate

Σ	2	dbo:background | group_or_band (236);
		dbo:activeYearsEndYear | 2010 (236);
			238	15	93,70%

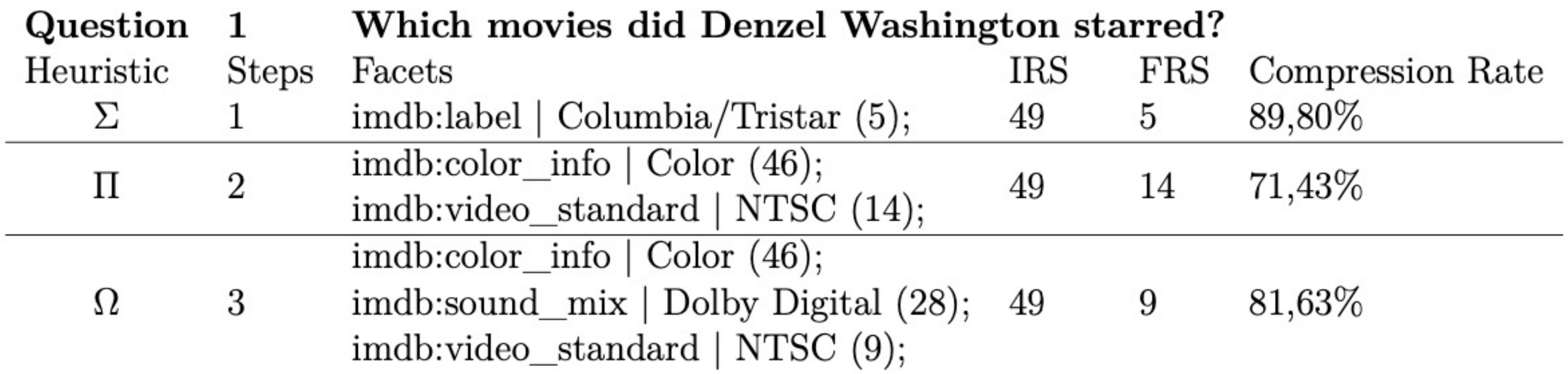
Π	2	dbo:background | group_or_band (236);
		dbo:activeYearsEndYear | 2010 (236);
			238	15	93,70%

Ω	2	dbo:activeYearsEndYear | 2010 (238);
		dbo:background | group_or_band (236);			

- In all cases, the process selected meaningful predicates but useless
  - selected predicates were related to the original question
  - group_or_band and 2010 compressed almost nothing
- The *Compression Rate* was good because of β threshold

Experiments

Compression Rate and final analysis - IMDb



Question	**1**	**Which movies did Denzel Washington starred?**			
Heuristic	Steps	Facets	IRS	FRS	Compression Rate
Σ	1	imdb:label | Columbia/Tristar (5);	49	5	89,80%

Π	2	imdb:color_info | Color (46);
		imdb:video_standard | NTSC (14);
			49	14	71,43%

Ω	3	imdb:color_info | Color (46);
		imdb:sound_mix | Dolby Digital (28);
		imdb:video_standard | NTSC (9);
			49	9	81,63%

- Except the heuristic Σ, the other heuristics in this first question selected some unhelpful predicates/facets
- Analyzing the other *Compression Rates*, all results were good because of β threshold
  - all other questions had no restrict facet and the original result set was long
  - κ <<< η since κ = β = 15

Experiments

Compression Rate and final analysis - IMDb

- We decided to investigate further the lack of predicates with restrictive facets in this dataset
- The questions evaluated were related to resources:
  - `imdb:Movie`
  - `imdb:Actor` and `imdb:Actress` – called *Artists*

Experiments

Compression Rate and final analysis - IMDb

#	Predicate	Distinct Values
1	rdfs:label	512491
2	imdb:title	512469
3	imdb:release_dates	274127
…	…	…
53	imdb:subtitles	12
54	imdb:analog_left	8
55	imdb:picture_format	8
56	imdb:sound_encoding	8
57	imdb:digital_sound	7
58	imdb:analog_right	6
59	imdb:category	4
60	imdb:color_information	4
61	imdb:status_of_availablility	4
62	imdb:video_standard	4
63	imdb:close_captions-teletext-ld-g	3
64	imdb:disc_format	3
65	imdb:disc_size	3
66	imdb:master_format	3
67	imdb:color_info	2
68	imdb:quality_program	1

`imdb:Movie` results

#	Predicates	Distinct Values
1	imdb:name	2192378
2	rdfs:label	2191943
3	imdb:trivia	495432
4	imdb:aka	470130
5	imdb:other_works	305969
…	…	…
22	imdb:salary_history	5384
23	imdb:biographical_movies	4648
24	imdb:portrayed_in	3585
25	imdb:height	444
26	imdb:gender	2

`Artists'` results

- The set of selectable predicates is small, and their meaningfulness is also low
- Except `imdb:category` and `imdb:gender` all predicates are related to technical information about the movies

Conclusion and Future Work

- The main contribution of this study was the definition of a process - called *Query Answer Modification Process*
  - addressed the problem related to open-ended questions (long result sets)
  - simplicity is key (combined heuristics with set of thresholds)
- Results compared and discussed over known baselines
  - QALD challenge and Coffman's benchmark
  - *Compression Rate* metric define to compare results

Conclusion and Future Work

- The initial research involving the *Query Answer Modification Process* was published in the Proceedings of the XXXV Brazilian Symposium on Databases – SBBD
- An extended version of this paper was submitted to the Journal of Information and Data Management (JIDM) and was approved
- Suggestions for improvements - part 1:
  - allow users to provide a list of ignored predicates
    - or enable dynamically exclusion of undesired predicates in each step
  - develop and apply a questionnaire to users interested in using the system
    - we would be able to segment users by preferences and create user profiles
    - the system would be able to adapt heuristics based on associated profile

Conclusion and Future Work

- Suggestions for improvements - part 2:
  - allow users to inform multi-level of specific predicates (taxonomy)
    - or automatically infer the related taxonomy during initial exploration of the RDF graph
  - develop a user interface similar to GraFa
    - navigation through predicates and facets
    - users of all types (beginners or experienced) would be able to use and evaluate our process
  - allow users to select more than one option at each interaction
    - disjunctive filtering would be possible, together with conjunctive filtering already developed

Thank you!

João Pedro V. Pinheiro[1], Marco A. Casanova[1], Elisa S. Menendez[2]

[1]Department of Informatics, PUC-Rio, Rio de Janeiro, RJ, Brazil

[2]Federal Institute of Education, Science and Technology of Bahia, Xique-Xique, BA, Brazil

{jpinheiro,casanova}@inf.puc-rio.br, elisa.menendez@ifbaiano.edu.br

http://www.inf.puc-rio.br/~jpinheiro/