ANT Searching for Information at the University of Porto

José Devezas
joseluisdevezas@gmail.com

INESC TEC & FEUP InfoLab
MAP-i 2016/2017

Universidade do Minho

universidade de aveiro

U.PORTO

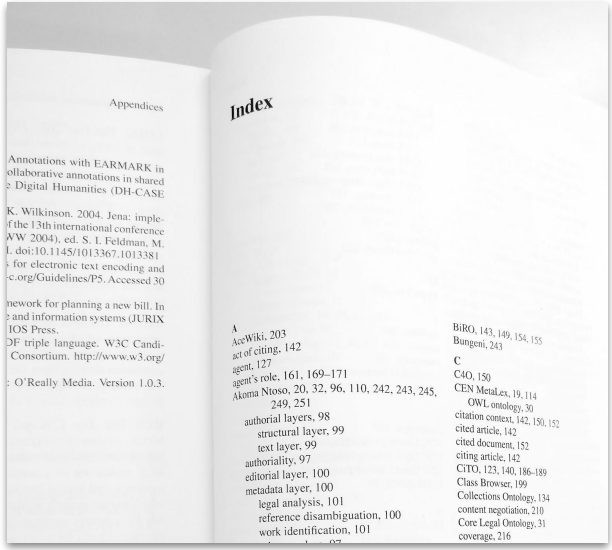
Contents

Introduction

What is entity-oriented search and why does it matter?

Search using keyword-based matching

- Modeled after the back-of-the-book index.
- Finding relevant content involves:
  1. Selecting one or several keywords;
  2. Jumping to the indicated pages;
  3. Reading passages and using knowledge, either internal or external to the book, to assess the relevance.



Appendices

Annotations with EARMARK in
ollaborative annotations in shared
e Digital Humanities (DH-CASE

K. Wilkinson. 2004. Jena: imple-
f the 13th international conference
AW 2004), ed. S. I. Feldman, M.
I. doi:10.1145/1013367.1013381
for electronic text encoding and
c.org/Guidelines/P5. Accessed 30

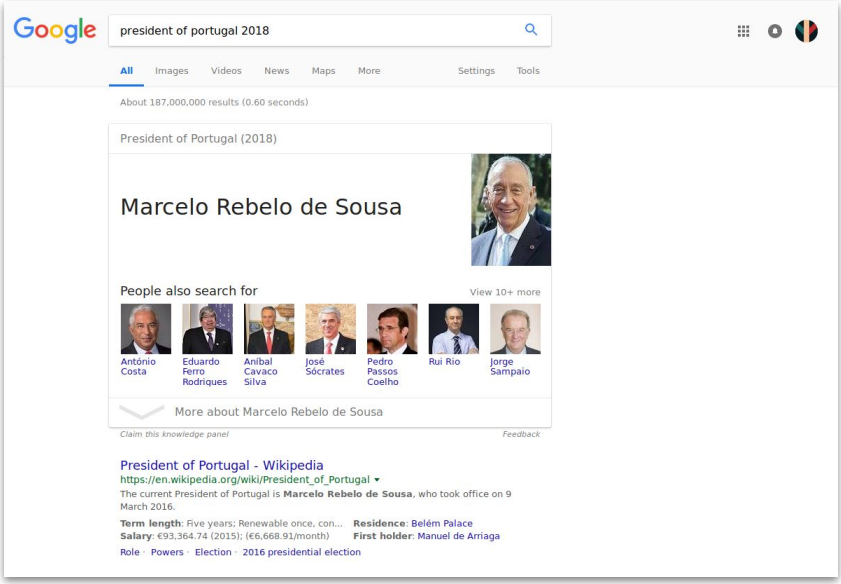
nework for planning a new bill. In
e and information systems (JURIX
IOS Press.
DF triple language. W3C Candi-
Consortium. http://www.w3.org/

: O'Really Media. Version 1.0.3.

Index

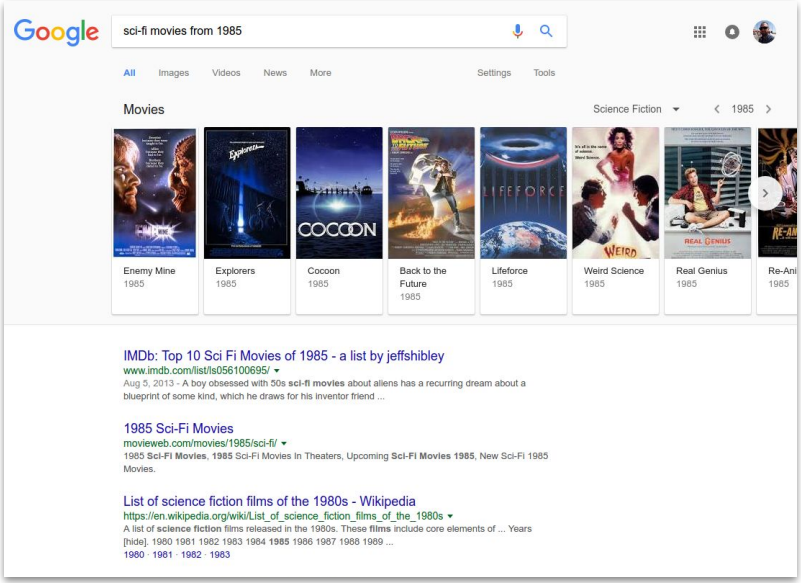
Search using semantic matching

- Closer the user's information need.
- Requires interpretation of query meaning and document semantics.
- Combines unstructured data from text and structured data from knowledge bases.
  - Google Knowledge Graph
  - Google Knowledge Vault
  - DBpedia
  - Wikidata



Search using semantic matching

- It becomes possible to, more adequately, answer queries like:
  - `[ president of portugal 2018 ]`
  - `[ sci-fi movies from 1985 ]`
- Instead of documents containing these keywords;
- An entity or list of entities is directly provided as the answer.
- Avoids the hassle of having to skim through documents to find the answer.



The relevance of entities in search

In queries:

- A 2007 study of the AOL Query Log showed that:
  - 18-39% queries directly refer to entities;
  - 73-87% queries contain at least one entity.

In documents:

- The annotated CoNLL 2003 English training set contains:
  - 14,987 sentences;
  - 23,499 entities;
  - 1.6 entities per sentence.

Entity-oriented search

Depending on the query, results can be:

- Documents
  - Retrieved using semantic information (entities and their relations).
- Entities
  - A specific one, a list, or both.
  - Retrieved by name, type, or another description.
  - Representing attributes or relations.

In ANT, entities can be:

- Students
- Staff
- Departments
- Rooms
- Curricular Units
- Courses
- News

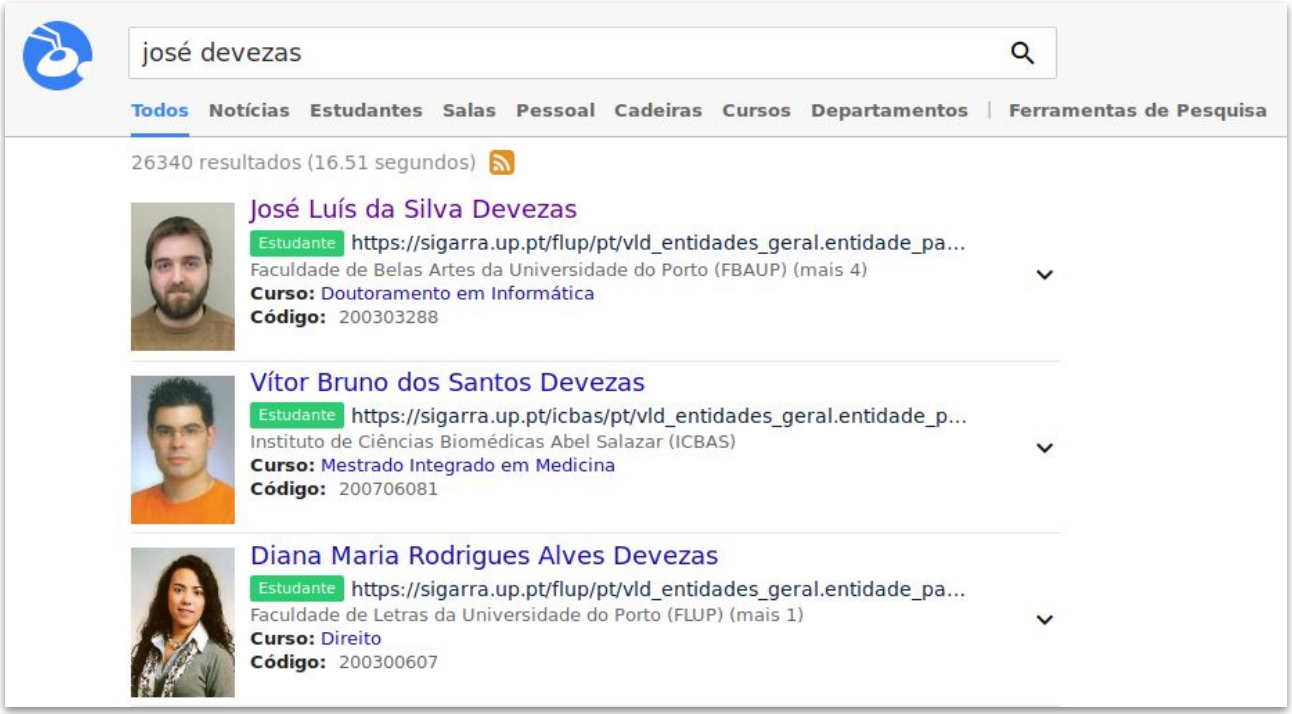
ANT

Search engine architecture and query understanding method.

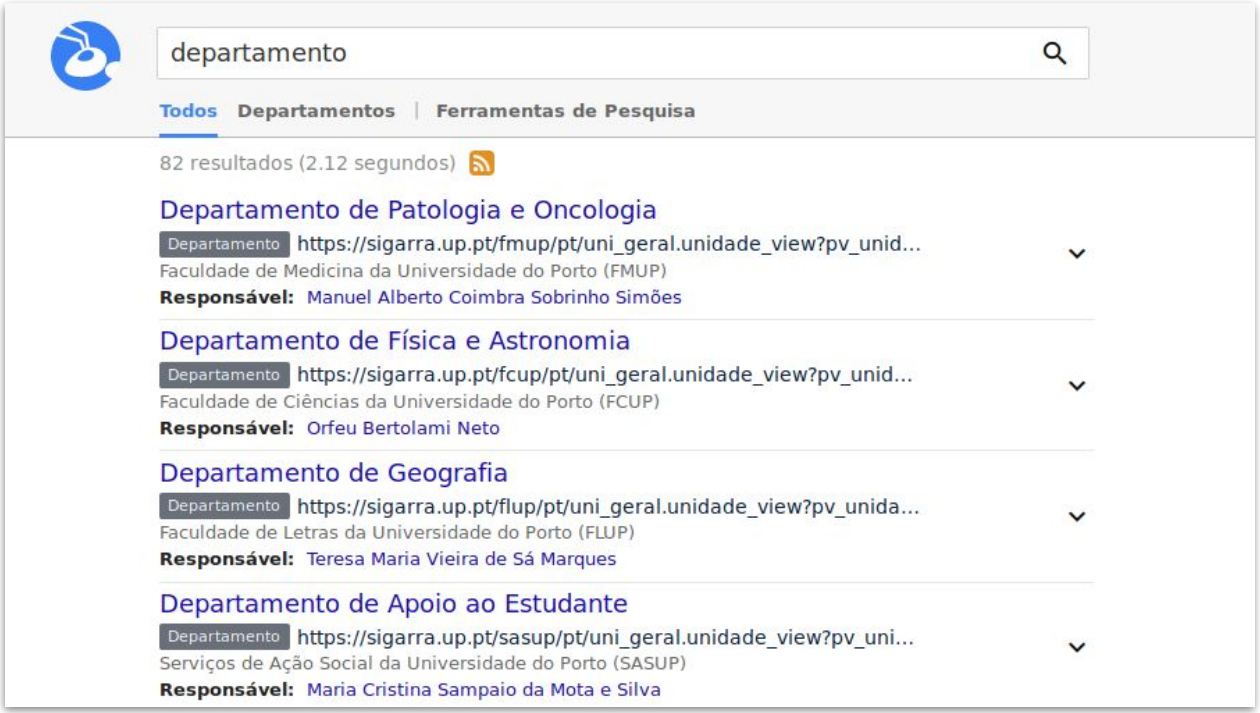
ANT

Ad hoc search of eNtities and Text.

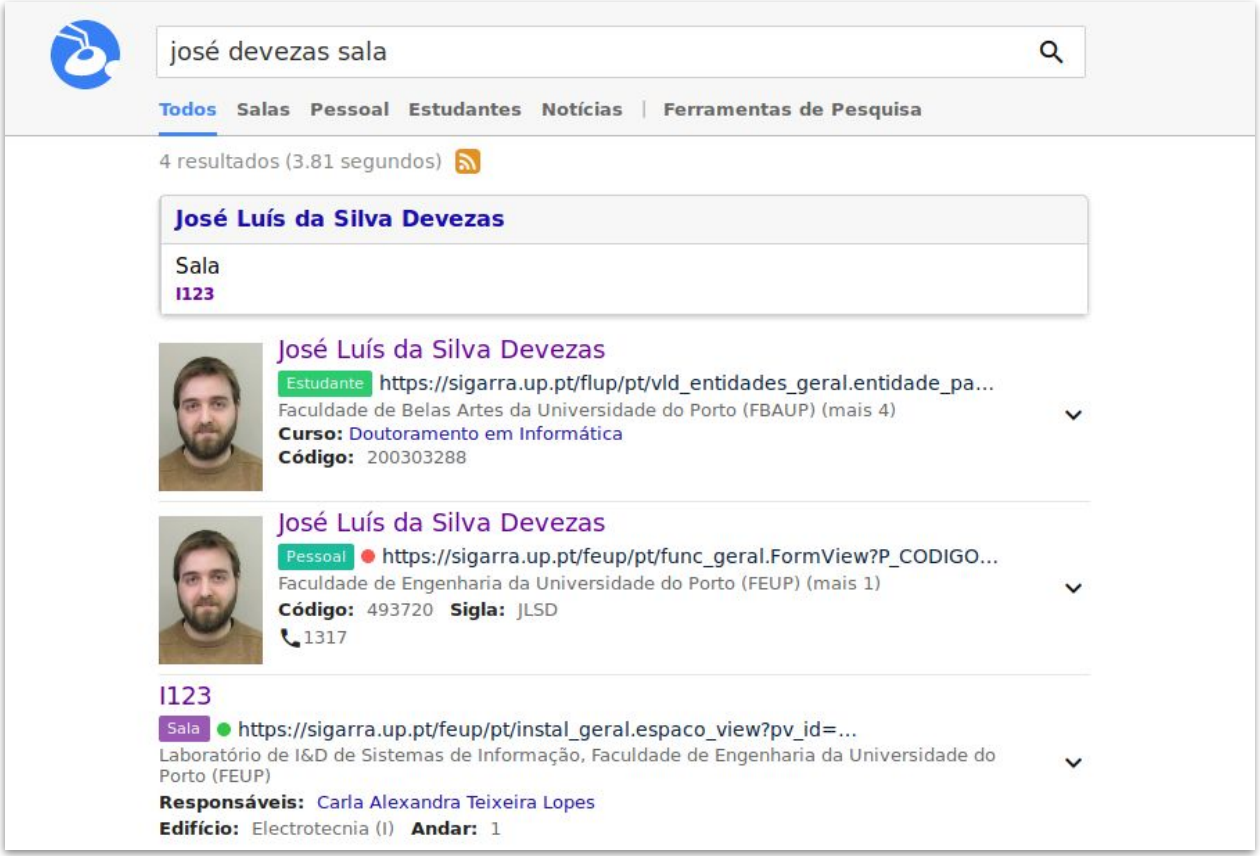
- Supports the five query categories defined by Pound et al. (2010):
  - Entity;
  - Type;
  - Attribute;
  - Relation;
  - Keyword.
- Based on two Lucene indexes:
  - Query analysis index;
  - Entity index.
- And a Virtuoso quad store:
  - Userful for attribute and relation queries.



Entity query. *The intention of the query is to find an entity.*



Type query. *The intention of the query is to find entities of a given type or class.*



Attribute query. *The intention of the query is to find values for a given attribute of a particular entity or type.*

ANT Pesquisa Orientada a Entidades Início Sobre Colaboração Contactos

josé devezas tiago devezas Pesquisar ?

6 resultados obtidos em 2,88 segundos

José Luís da Silva Devezas
Faculdade de Belas Artes da Universidade do Porto (FBAUP), Faculdade de Letras da Universidade do Porto (FLUP), Faculdade de Engenharia da Universidade do Porto (FEUP), Faculdade de Economia da Universidade do Porto (FEP), Faculdade de Ciências da Universidade do Porto (FCUP)
Estudante
score = 0.002238

Tiago Nuno Mesquita Folgado leitão Devezas
Faculdade de Engenharia da Universidade do Porto (FEUP) Domingo, 07 Outubro 2018, 00h32
Pessoal
score = 0.001971

José Luís da Silva Devezas
Faculdade de Engenharia da Universidade do Porto (FEUP), Faculdade de Ciências da Universidade do Porto (FCUP) Domingo, 02 Setembro 2018, 00h26
Pessoal
score = 0.001106

Tiago Nuno Mesquita Folgado Leitão Devezas
Faculdade de Belas Artes da Universidade do Porto (FBAUP), Faculdade de Letras da Universidade do Porto (FLUP), Faculdade de Engenharia da Universidade do Porto (FEUP), Faculdade de Economia da Universidade do Porto (FEP), Faculdade de Ciências da Universidade do Porto (FCUP)
Estudante
score = 0.001789

I123
Faculdade de Engenharia da Universidade do Porto (FEUP)
Sala
score = 0.000431

Projetos no laboratório SAPO/U.Porto
Faculdade de Engenharia da Universidade do Porto (FEUP) Sexta-feira, 28 Junho 2013, 00h00
Notícia
score = 0.0

<< Anterior Página 1 de 1 Seguinte >>

Entidades

Tiago Nuno Mesquita Folgado leitão Devezas

José Luís da Silva Devezas

Ligações

Faculdade: Faculdade de Engenharia, Faculdade de Engenharia da Universidade do Porto

Sala: I123

Projetor	Não
Edifício	Electrotecnia (I)
Mapa	https://sigarra.up.pt/feup/pt/instal_geral2.get_mapa?pv_id=77486

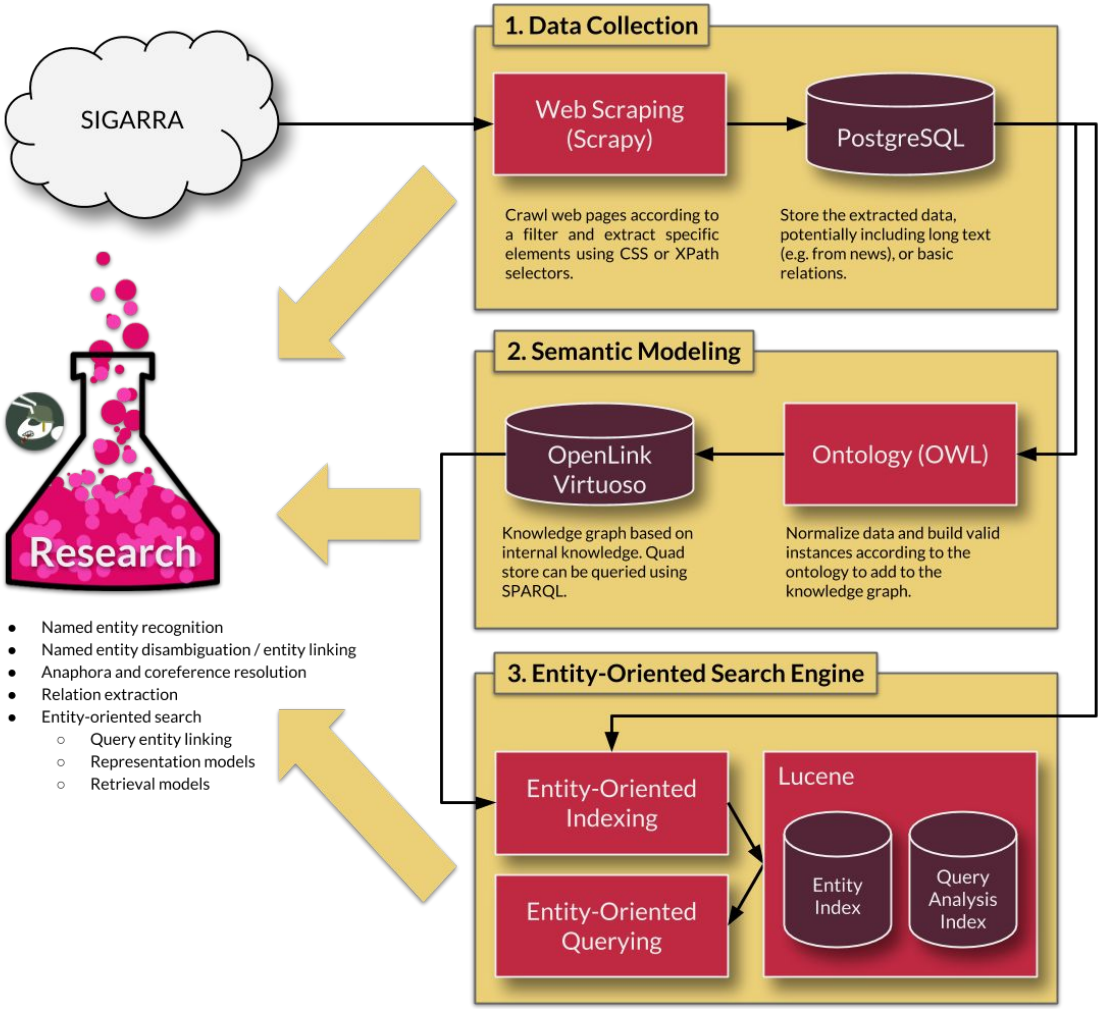
Ver mais

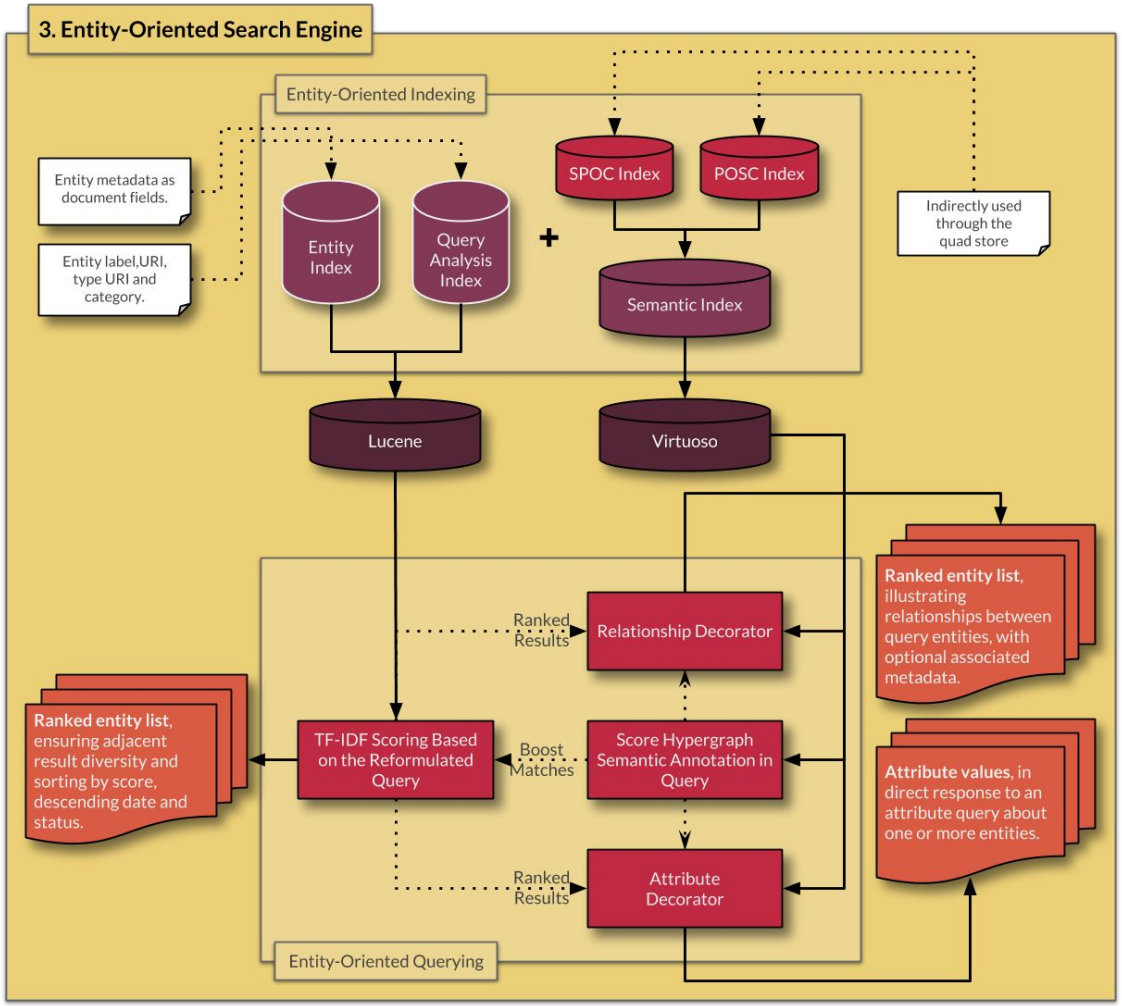
Relation query. *The intention of the query is to find how two or more entities or types are related.*

Keyword query. *Anything that doesn't fit the previous four categories.*

Search engine architecture

ANT components, from data collection to search.


SIGARRA
1. Data Collection
Web Scraping (Scrapy)
PostgreSQL
Crawl web pages according to a filter and extract specific elements using CSS or XPath selectors.
Store the extracted data, potentially including long text (e.g. from news), or basic relations.
2. Semantic Modeling
OpenLink Virtuoso
Ontology (OWL)
Knowledge graph based on internal knowledge. Quad store can be queried using SPARQL.
Normalize data and build valid instances according to the ontology to add to the knowledge graph.
3. Entity-Oriented Search Engine
Entity-Oriented Indexing
Entity-Oriented Querying
Lucene
Entity Index
Query Analysis Index
Research
Named entity recognition
Named entity disambiguation / entity linking
Anaphora and coreference resolution
Relation extraction
Entity-oriented search
Query entity linking
Representation models
Retrieval models


3. Entity-Oriented Search Engine
Entity-Oriented Indexing
Entity metadata as document fields.
Entity label,URI, type URI and category.
Entity Index
Query Analysis Index
+
SPOC Index
POSC Index
Semantic Index
Indirectly used through the quad store
Lucene
Virtuoso
Ranked Results
Relationship Decorator
Ranked entity list, illustrating relationships between query entities, with optional associated metadata.
Ranked entity list, ensuring adjacent result diversity and sorting by score, descending date and status.
TF-IDF Scoring Based on the Reformulated Query
Boost Matches
Score Hypergraph Semantic Annotation in Query
Attribute values, in direct response to an attribute query about one or more entities.
Ranked Results
Attribute Decorator
Entity-Oriented Querying

REST API

- ANT provides access to search-related services via a REST API.
- We use the OpenAPI 2.0 format (Swagger) to document the API.
  - http://ant.fe.up.pt/api/
  - https://swagger.io/specification/
- Which makes it possible to easily provide a console for API exploration.
  - http://ant.fe.up.pt/api-console/
- Supported services are classified into six categories:
  - Analytics
  - Autocomplete
  - Decorators*
  - JavaScript
  - Log
  - Search*

* Most relevant services.

Query understanding

What does it mean “to understand” and what is ANT’s approach?

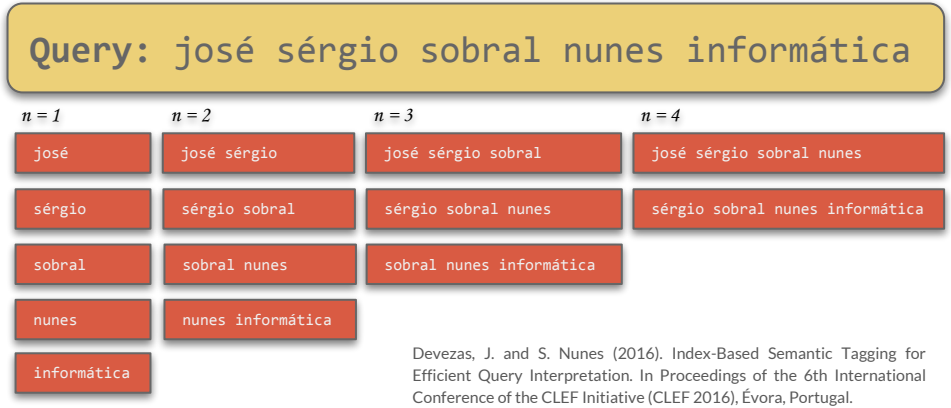
Query understanding

- Segment the query.
  - According to knowledge from the indexed corpus and from the knowledge base, which is the most likely division of the query?
- Annotate with semantic tags.
  - For each segment, assign the most likely semantic class (or none).
- Classify the query.
  - Based on the annotated segments, label the query with one of the five categories from Pound et al. (2010).

- In ANT, the query is seen as a sequence of keywords.
- But it could instead be considered as natural language:
  - Well-formed sentences;
  - Well-structured questions.
- A search engine's query processing methods should vary accordingly, with the style of query or, even, the user interface.

How does ANT understand queries?

- Query segmentation based on the retrieval of matching entities for all query n-grams up to a maximum value of n.
- Semantic tagging of query segments based on the probability of associating a given type of entity to an n-gram.



Devezas, J. and S. Nunes (2016). Index-Based Semantic Tagging for Efficient Query Interpretation. In Proceedings of the 6th International Conference of the CLEF Initiative (CLEF 2016), Évora, Portugal.

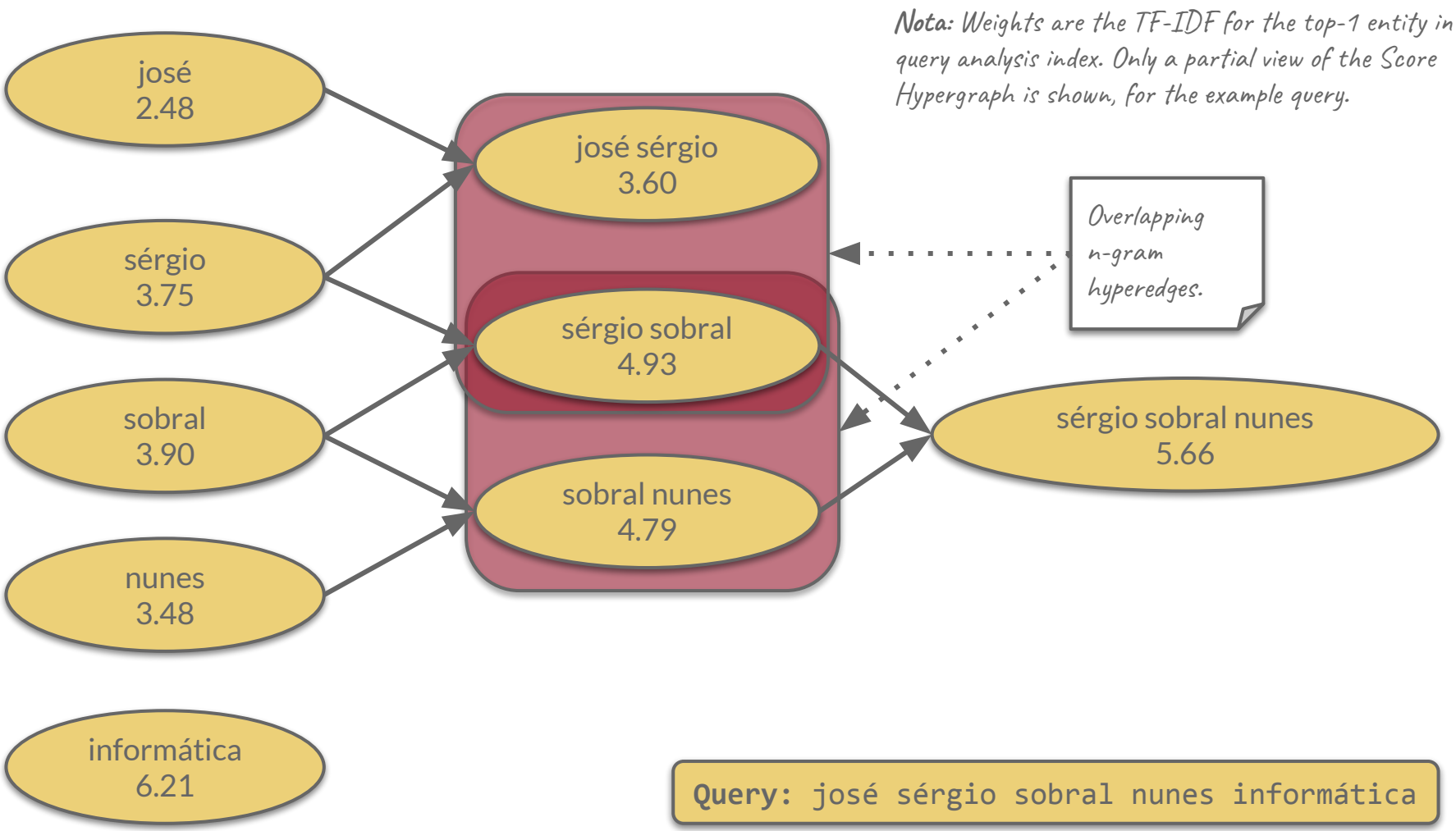
How does ANT understand queries?

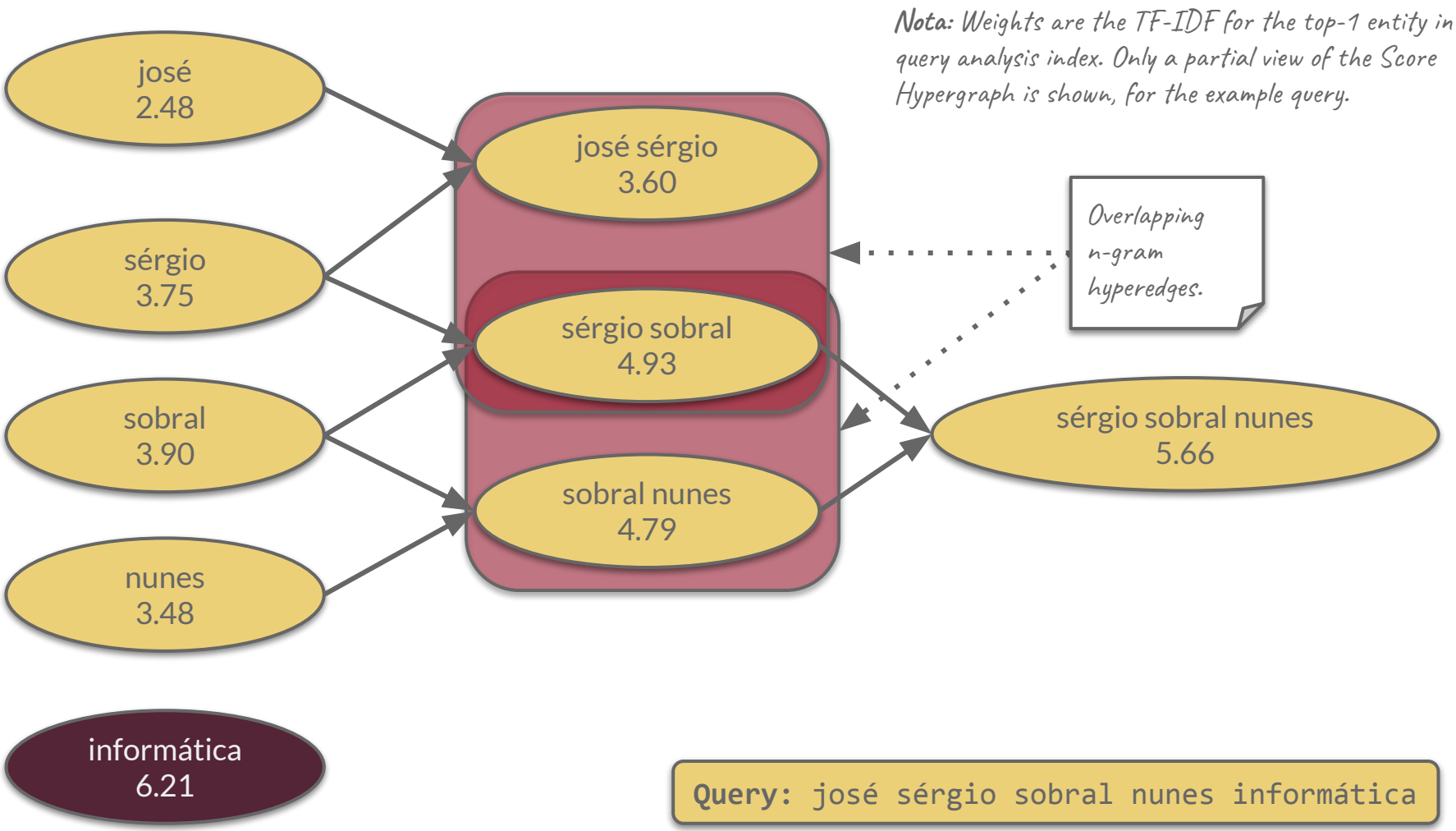
- In reality, we ended up using "**Score Hypergraph**" instead, which is a slight variation of the previous method, improving on performance and fixing some bugs.
- We used TF-IDF scores instead of probabilities.
- We created a dedicated index for query analysis, in order to search for entities matching n-grams, instead of querying the triple store.
- We used a hypergraph* of n-grams to solve query segment overlaps.

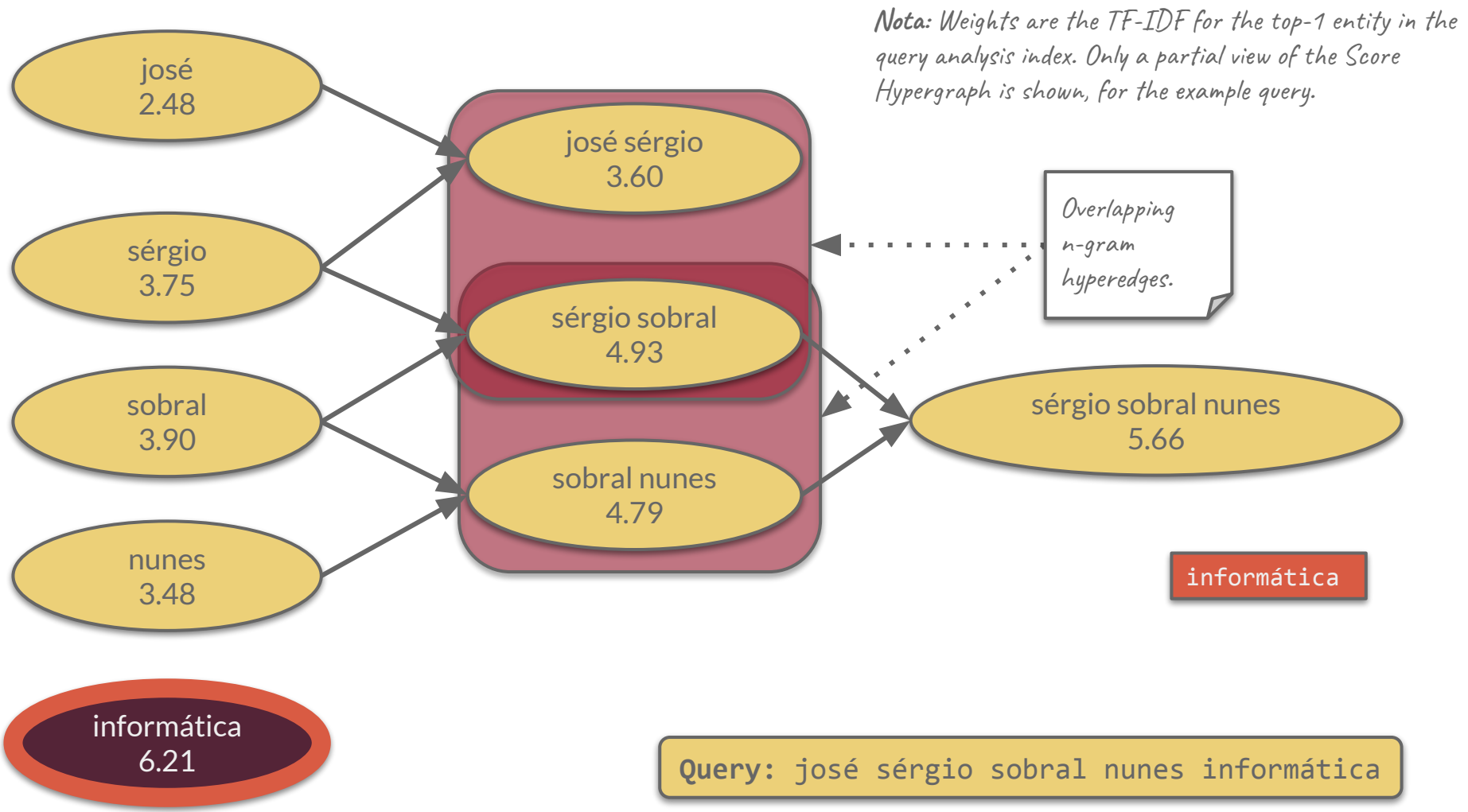
* A hypergraph is a generalization of a graph, where edges can have an arbitrary number of nodes.

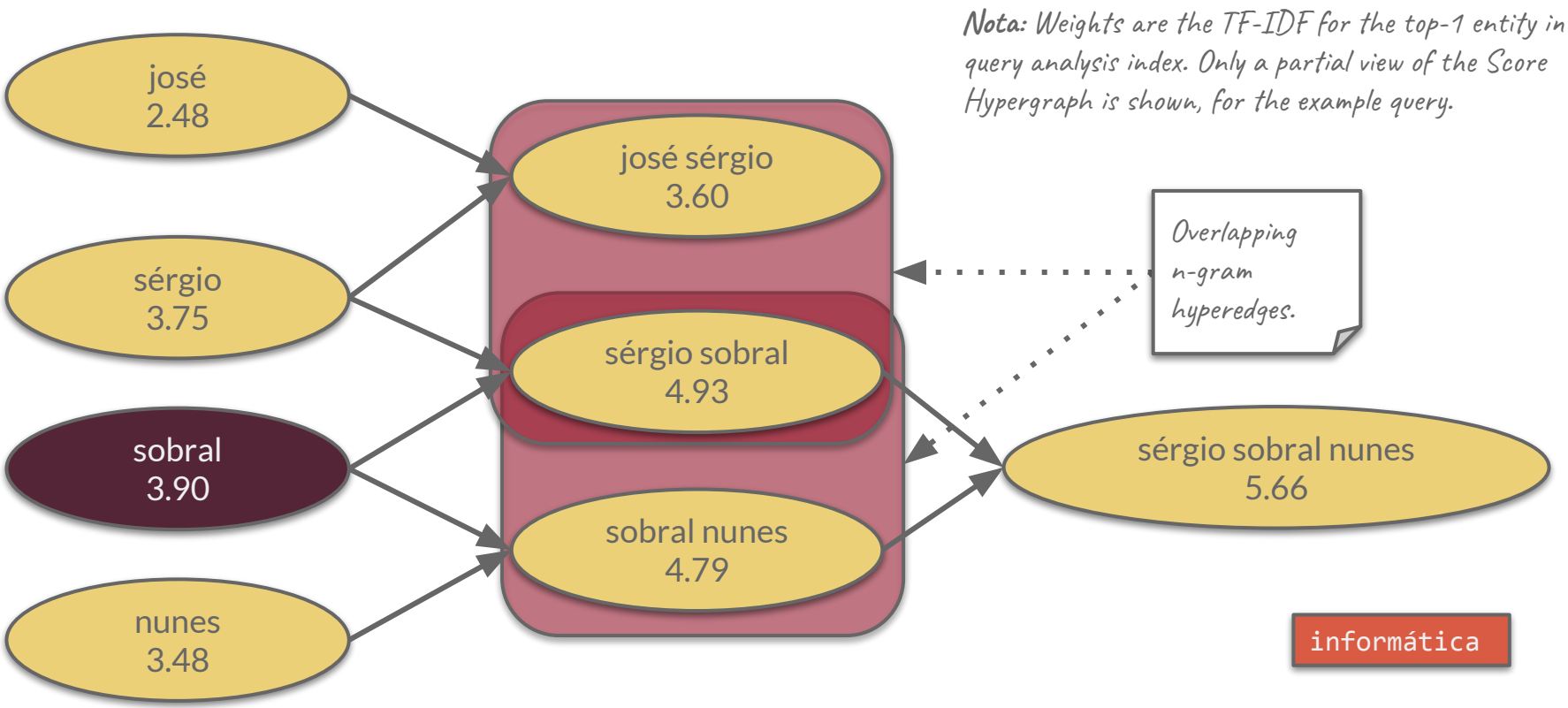
Score Hypergraph

Query segmentation and semantic tagging in ANT.

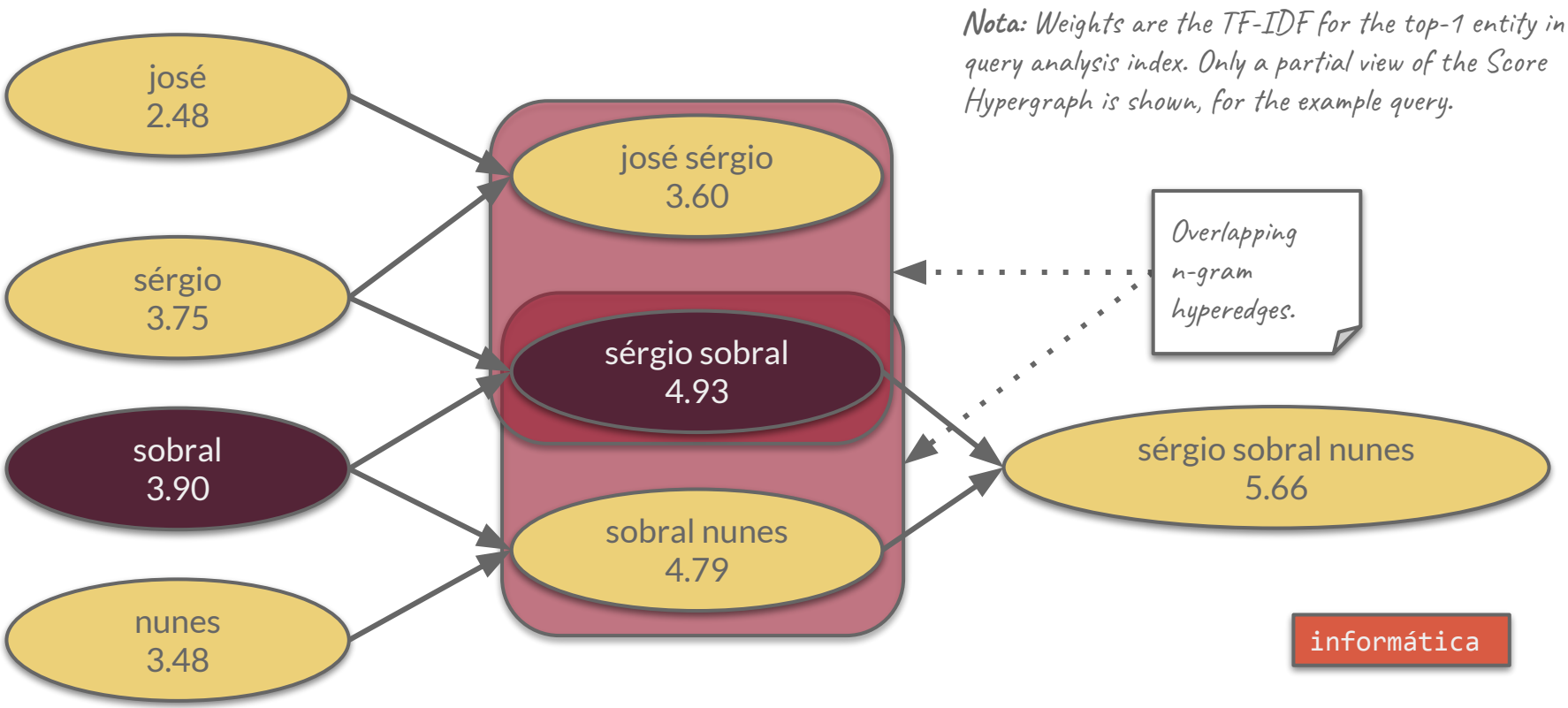

Nota: Weights are the TF-IDF for the top-1 entity in the query analysis index. Only a partial view of the Score Hypergraph is shown, for the example query.
josé
2.48
sérgio
3.75
sobral
3.90
nunes
3.48
informática
6.21
josé sérgio
3.60
sérgio sobral
4.93
sobral nunes
4.79
sérgio sobral nunes
5.66
Overlapping n-gram hyperedges.
Query: josé sérgio sobral nunes informática


Nota: Weights are the TF-IDF for the top-1 entity in the query analysis index. Only a partial view of the Score Hypergraph is shown, for the example query.
josé
2.48
sérgio
3.75
sobral
3.90
nunes
3.48
informática
6.21
josé sérgio
3.60
sérgio sobral
4.93
sobral nunes
4.79
sérgio sobral nunes
5.66
Overlapping n-gram hyperedges.
Query: josé sérgio sobral nunes informática

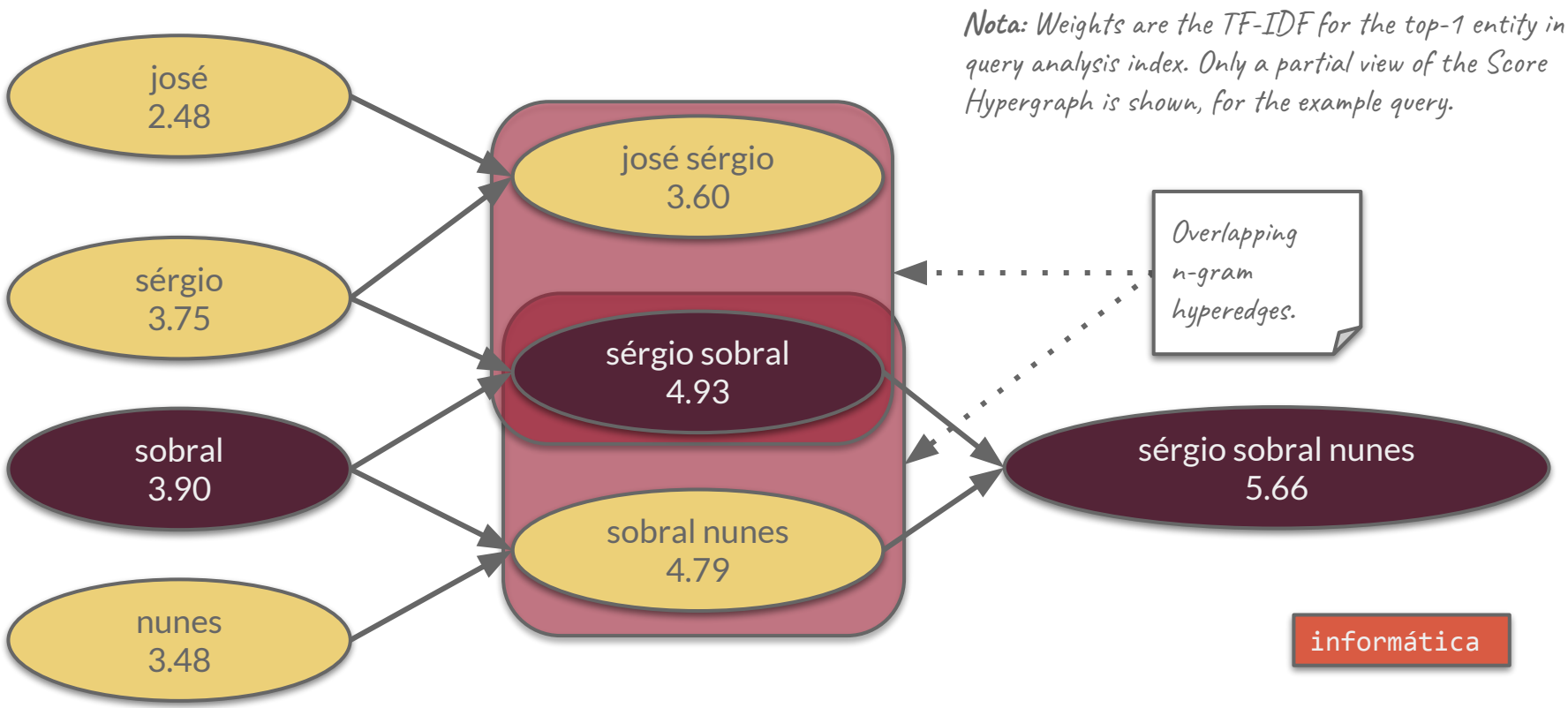

Nota: Weights are the TF-IDF for the top-1 entity in the query analysis index. Only a partial view of the Score Hypergraph is shown, for the example query.
josé
2.48
sérgio
3.75
sobral
3.90
nunes
3.48
josé sérgio
3.60
sérgio sobral
4.93
sobral nunes
4.79
Overlapping n-gram hyperedges.
sérgio sobral nunes
5.66
informática
informática
6.21
Query: josé sérgio sobral nunes informática



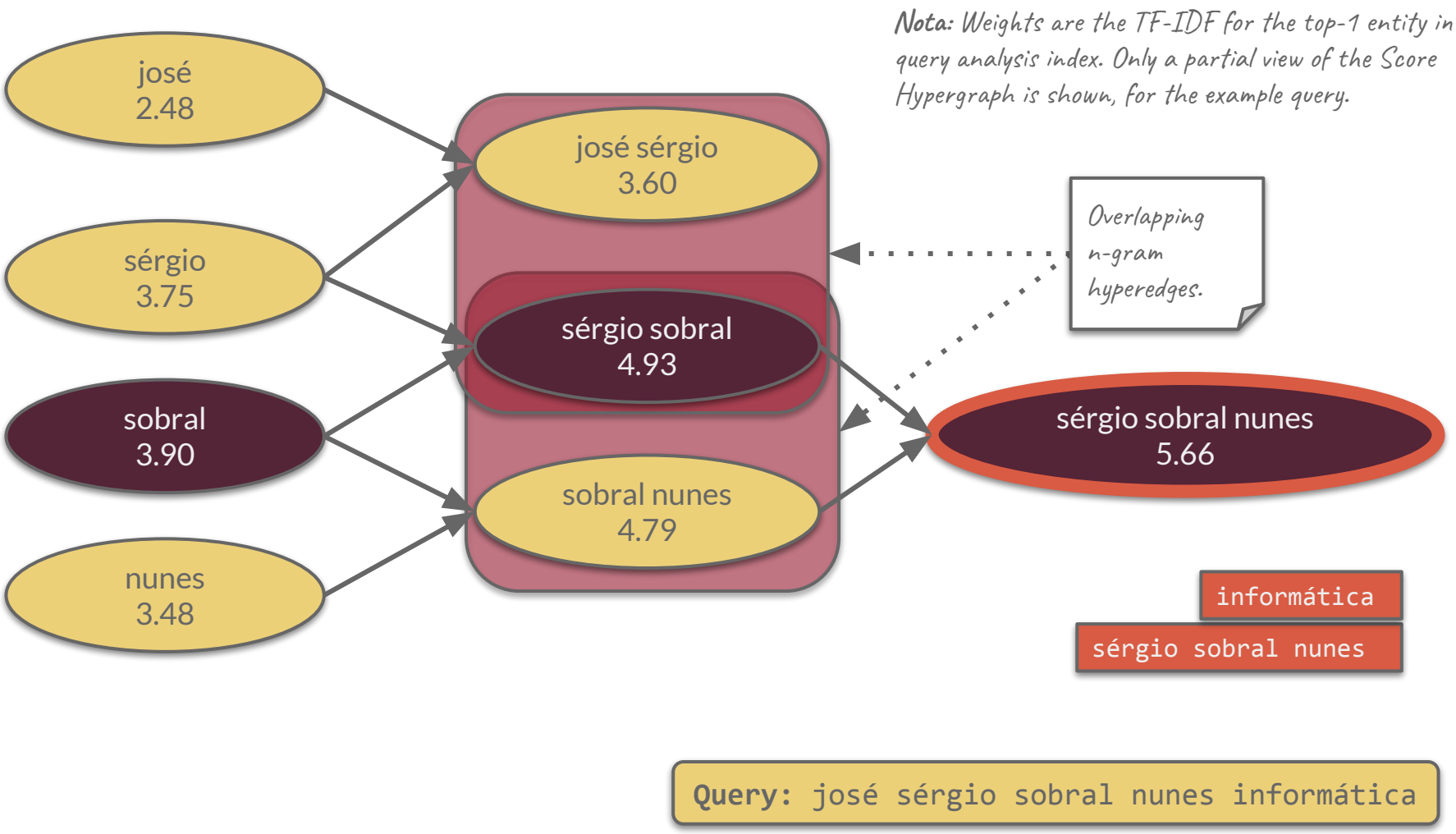
Query: josé sérgio sobral nunes informática

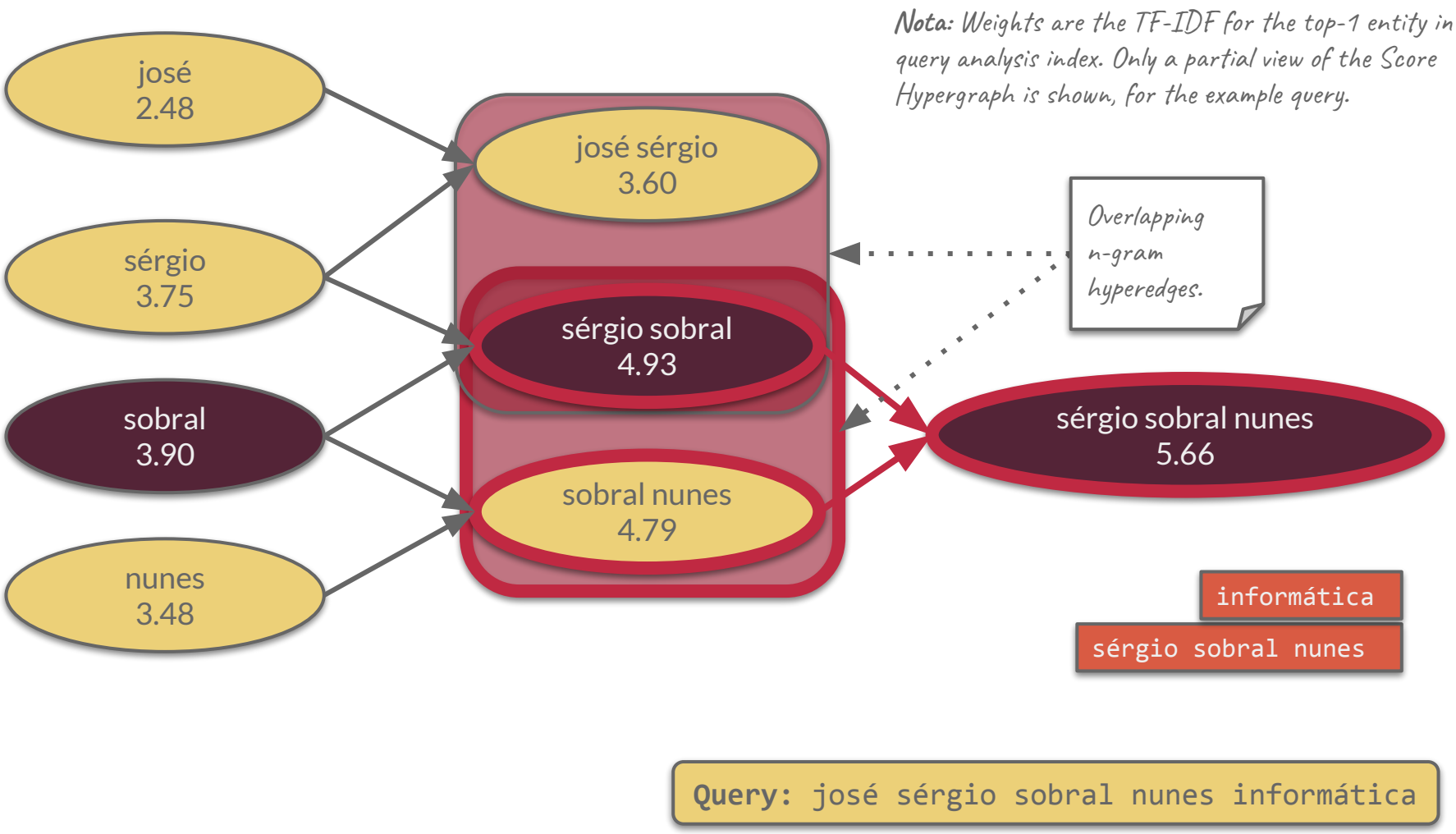


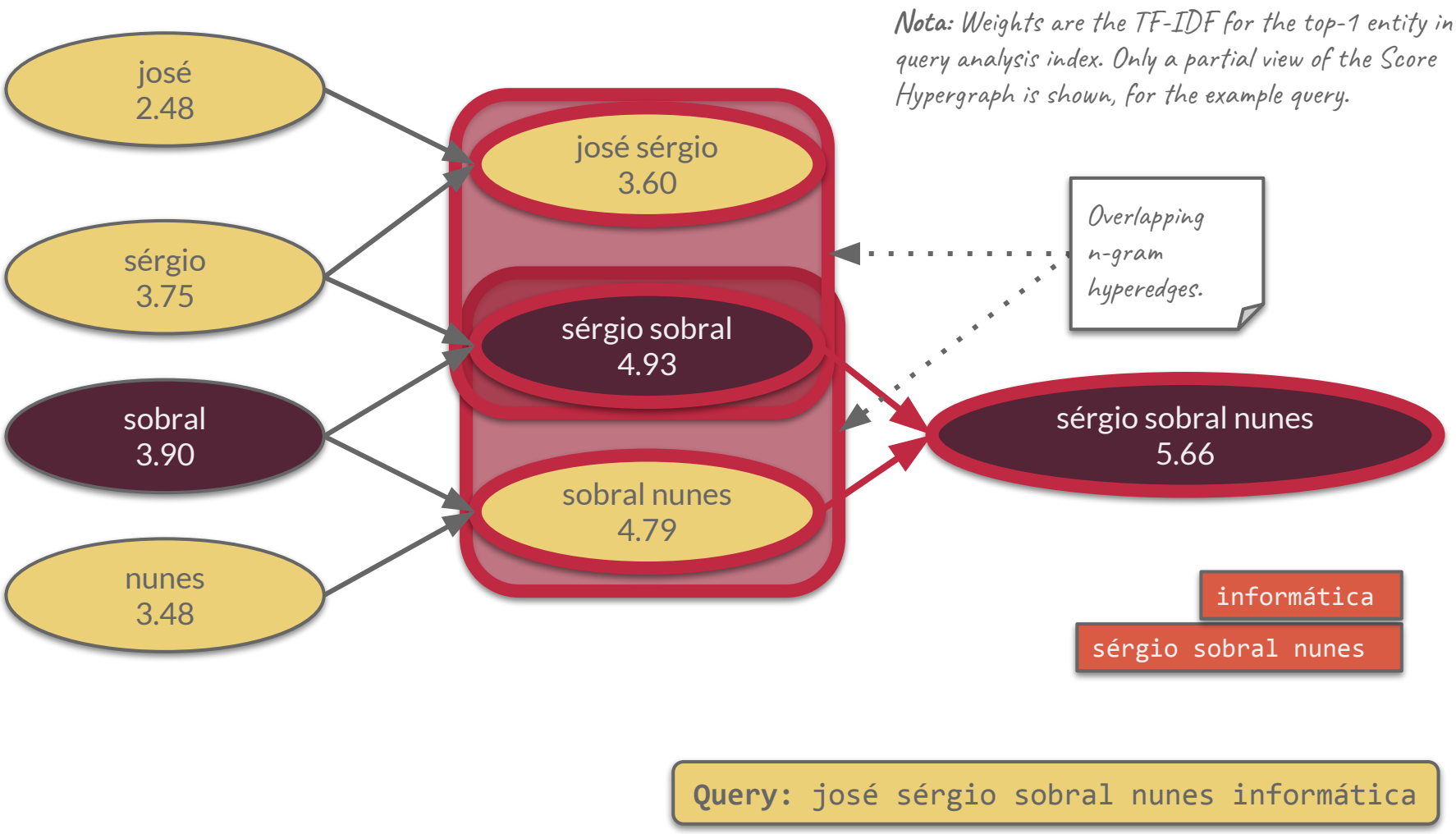
Query: josé sérgio sobral nunes informática

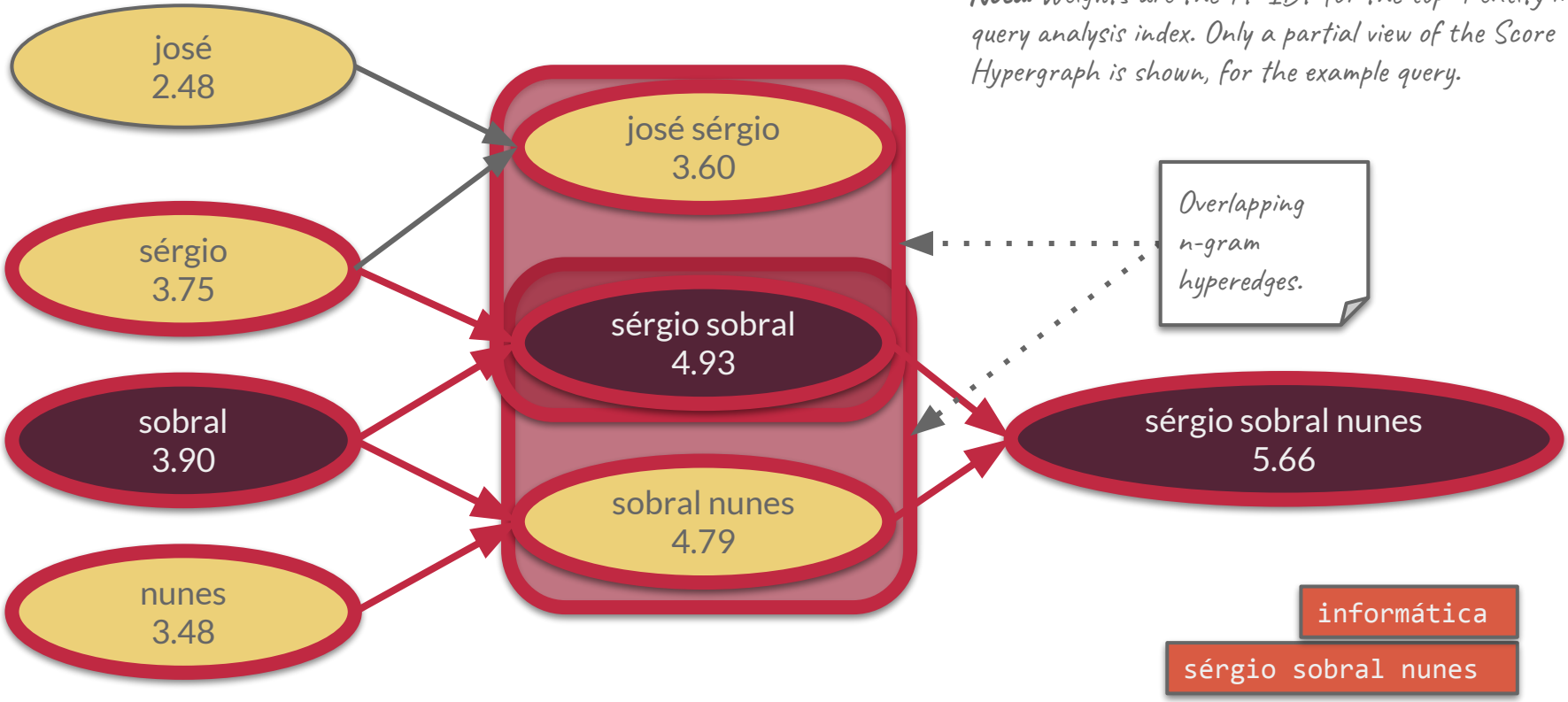


Query: josé sérgio sobral nunes informática


Nota: Weights are the TF-IDF for the top-1 entity in the query analysis index. Only a partial view of the Score Hypergraph is shown, for the example query.
josé
2.48
sérgio
3.75
sobral
3.90
nunes
3.48
josé sérgio
3.60
sérgio sobral
4.93
sobral nunes
4.79
Overlapping n-gram hyperedges.
sérgio sobral nunes
5.66
informática
sérgio sobral nunes
Query: josé sérgio sobral nunes informática


Nota: Weights are the TF-IDF for the top-1 entity in the query analysis index. Only a partial view of the Score Hypergraph is shown, for the example query.
josé
2.48
sérgio
3.75
sobral
3.90
nunes
3.48
josé sérgio
3.60
sérgio sobral
4.93
sobral nunes
4.79
Overlapping n-gram hyperedges.
sérgio sobral nunes
5.66
informática
sérgio sobral nunes
Query: josé sérgio sobral nunes informática


Nota: Weights are the TF-IDF for the top-1 entity in the query analysis index. Only a partial view of the Score Hypergraph is shown, for the example query.
josé
2.48
sérgio
3.75
sobral
3.90
nunes
3.48
josé sérgio
3.60
sérgio sobral
4.93
sobral nunes
4.79
Overlapping n-gram hyperedges.
sérgio sobral nunes
5.66
informática
sérgio sobral nunes
Query: josé sérgio sobral nunes informática



Nota: Weights are the TF-IDF for the top-1 entity in the query analysis index. Only a partial view of the Score Hypergraph is shown, for the example query.

Query: josé sérgio sobral nunes informática

josé
2.48

Nota: Weights are the TF-IDF for the top-1 entity in the query analysis index. Only a partial view of the Score Hypergraph is shown, for the example query.

informática

sérgio sobral nunes

Query: josé sérgio sobral nunes informática

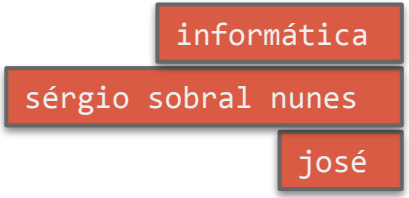
Nota: Weights are the TF-IDF for the top-1 entity in the query analysis index. Only a partial view of the Score Hypergraph is shown, for the example query.

informática

sérgio sobral nunes

Query: josé sérgio sobral nunes informática

Nota: Weights are the TF-IDF for the top-1 entity in the query analysis index. Only a partial view of the Score Hypergraph is shown, for the example query.



Query: josé sérgio sobral nunes informática

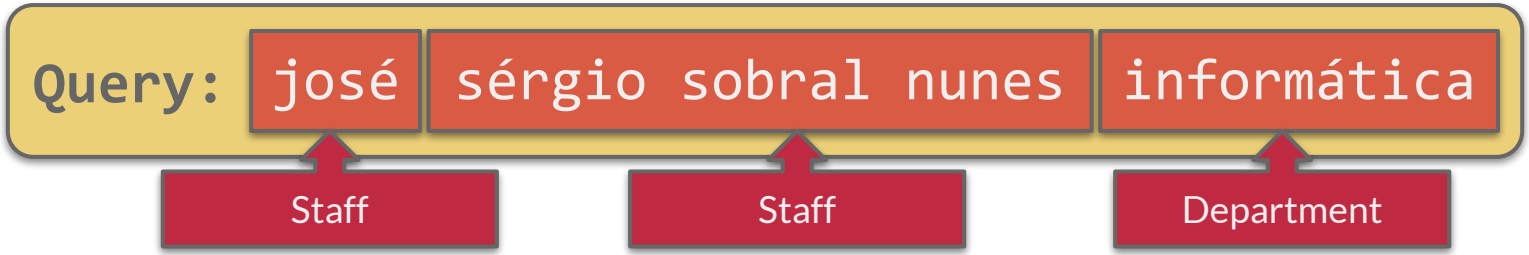
informática

sérgio sobral nunes

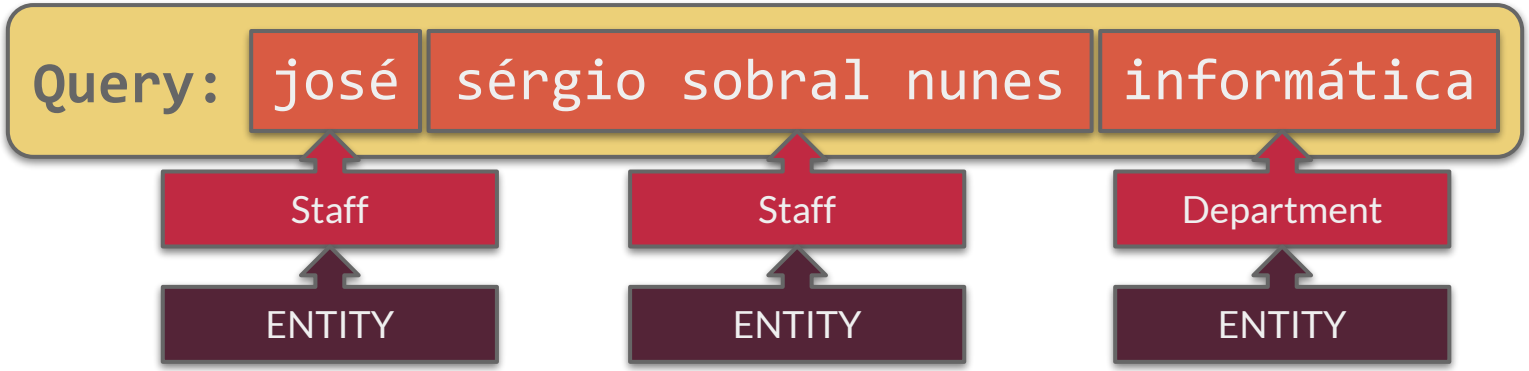
josé

Query: josé sérgio sobral nunes informática

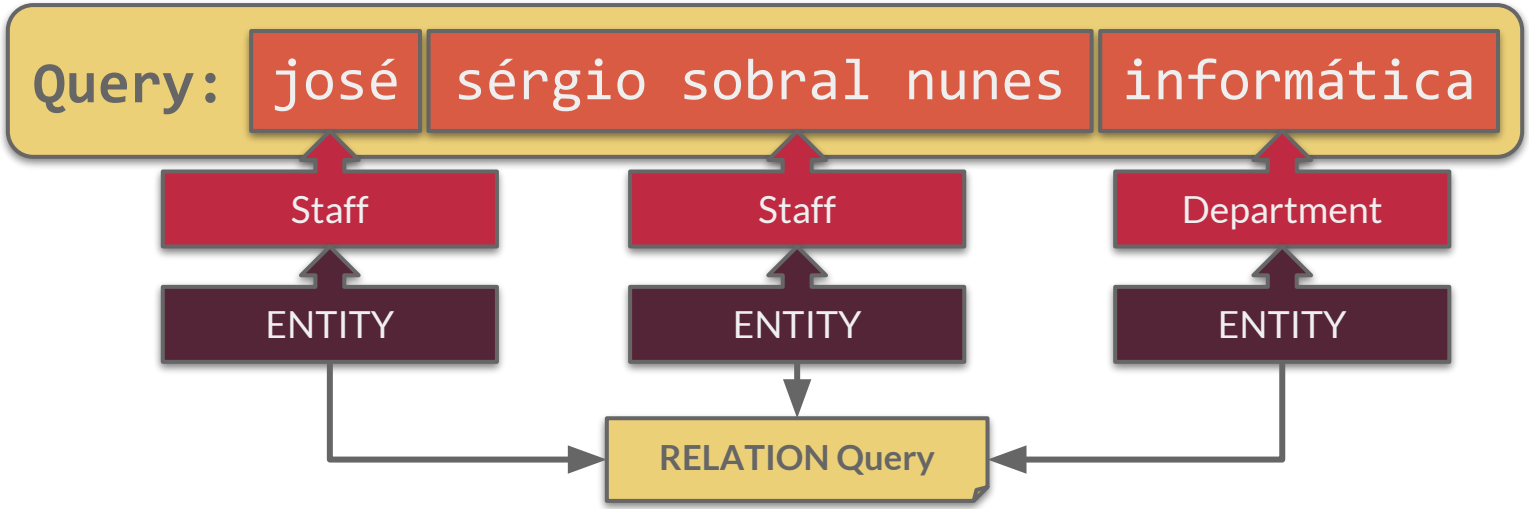
1. The query was segmented based on the n-grams with the highest-scoring entities.



2. The query was assigned semantic tags based on the type of the highest-scoring entity.



3. From the semantic tag, we directly derived a higher level tag that could either be ENTITY (e.g., instance of Staff class), ATTRIBUTE (e.g., property) or TYPE (e.g., Staff class).



4. Based on the combination of higher level tags, we conditionally obtained the query category.

Conclusions

Final remarks, related projects, and a vision for the future.

Final remarks

- The ANT search engine is serving the local academic community, despite infrastructure and human resource limitations (it's a prototype).
- At the same time, it collects implicit relevance feedback, based on result clicks for issued queries.
- ANT has also served as a platform of collaboration for multiple areas of research:
  - Web Design (collaboration with MM for the development of the front-end);
  - User Experience (MM dissertation in entity-oriented search interfaces);
  - Information Extraction (MIEIC dissertation in named entity recognition for portuguese web text).

Related projects

Army ANT

- Serving the research needs in the area of entity-oriented search.
- Supporting the study of innovative ideas in search, providing tools for exploration and evaluation.

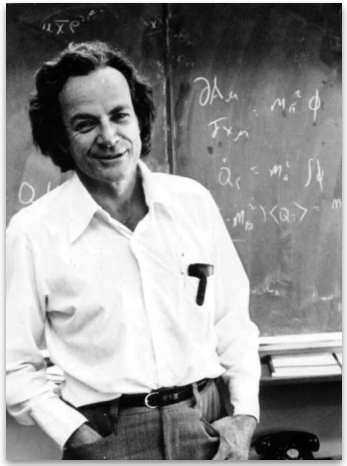
PhD thesis

- “Graph-Based Entity-Oriented Search”
  - Joint representation of text, entities and their relations.
  - Generalization of entity-oriented search tasks.
  - Improvement of search effectiveness?
- Exploration of random walks in graphs and hypergraphs.

A vision for the future...

...can easily begin in a partially forgotten past.

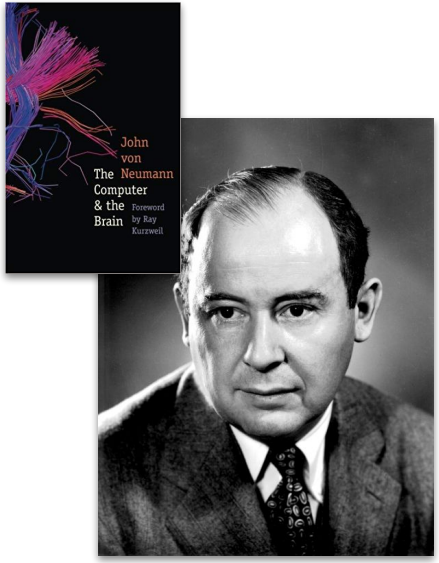
Von Neumann



Richard Feynman

A vision for the future...

...can easily begin in a partially forgotten past.



Von Neumann

Prior to his death, in 1957, John Von Neumann was preparing his classes for the Yale Silliman lectures. The idea was to compare the computer and the brain, studying both of their organs (neuron/transistor).

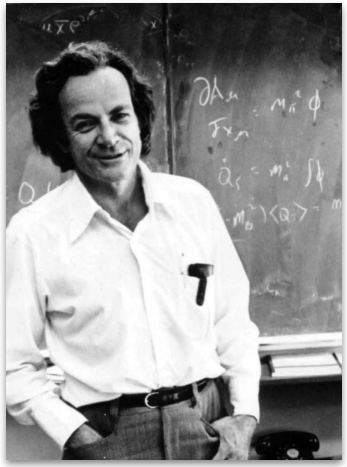
In the present, Cognitive Science already fits the role of tying together Neuroscience and Artificial Intelligence. Nevertheless, there are still unexplored paths at the intersections. There are answers and ideas that have been lost in the past!

A vision for the future...

...can easily begin in a partially forgotten past.

Theoretical physicist, with relevant contributions in Quantum Mechanics.

During World War II, he administered the group of human computers in the theoretical division of the Los Alamos Laboratory. Human computing is nothing else than manually solving a set of long and complex calculations, in group and strictly following pre-established rules.



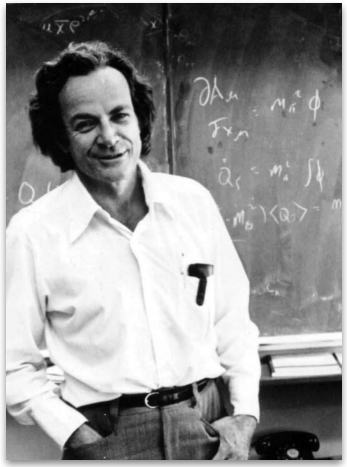
Richard Feynman

A vision for the future...

...can easily begin in a partially forgotten past.

Maybe that's why he collaborated with Stanley Frankel and Nicholas Metropolis in developing a system for using IBM punch cards for computation.

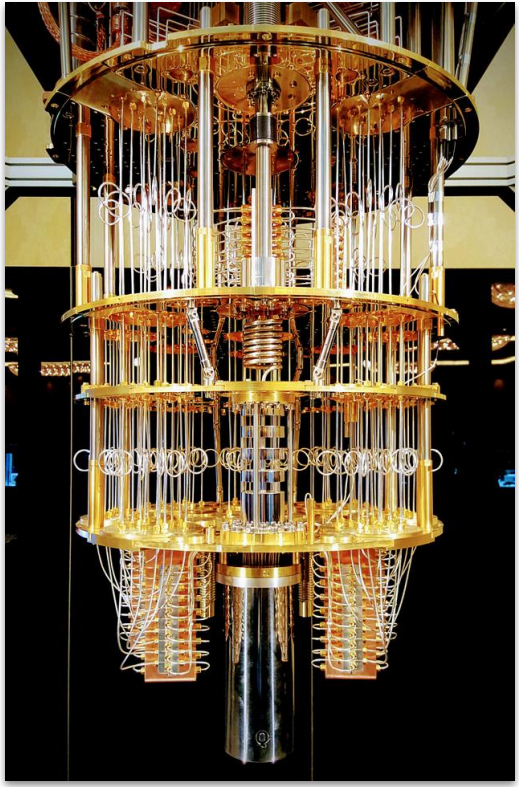
He also invented a new method for computing logarithms that he later used in the Connection Machine, developed by Daniel Hillis as an alternative to the Von Neumann architecture.



Richard Feynman

A vision for the future...

...towards quantum search engines?

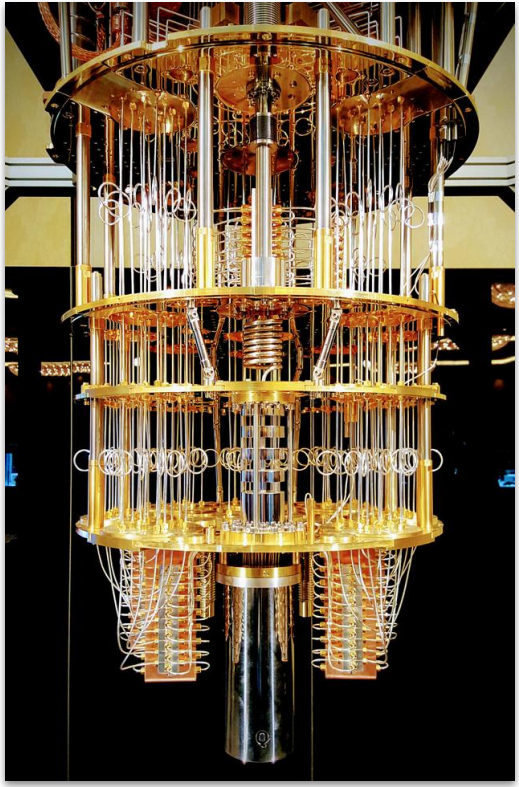


A vision for the future...

...with many unanswered questions.

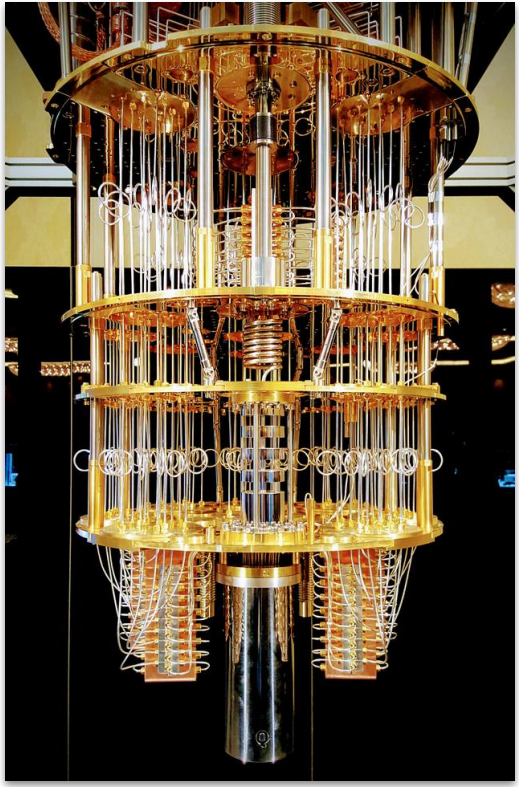
"Major Quantum Computing Advance Made Obsolete by Teenager"

https://www.quantamagazine.org/teenager-finds-classical-alternative-to-quantum-recommendation-algorithm-20180731/



A vision for the future...

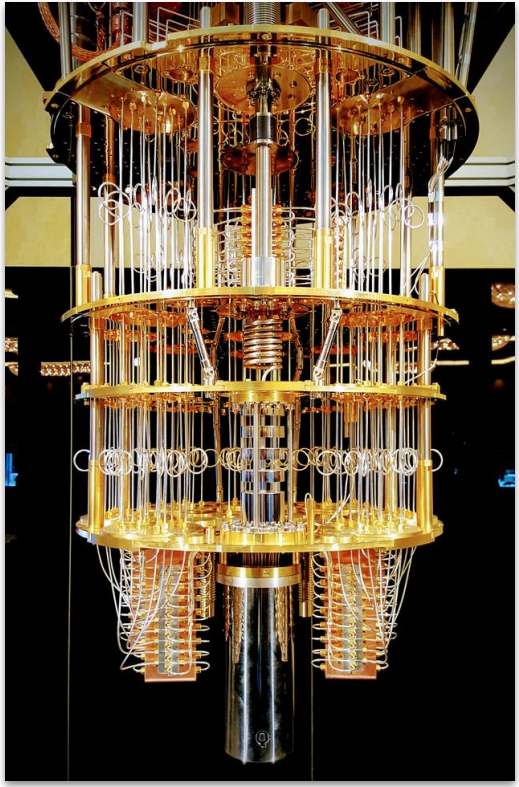
...where uncertainty and probabilities trades places with binary.



A vision for the future...

...where uncertainty-based search reinforces its relevance.

At FEUP InfoLab, we're working with classical computers, while exploring ideas that might have quantum applications.



A vision for the future...

...that is happier, more creative and increasingly exploratory.

The time is now, to combine the tools that science has already provided. The future is hybrid!

You mean... Electric?

Thank you!

https://ant.fe.up.pt

José Devezas is supported by research grant PD/BD/128160/2016, provided by the Portuguese national funding agency for science, research and technology, Fundação para a Ciência e a Tecnologia (FCT), within the scope of Operational Program Human Capital (POCH), supported by the European Social Fund and by national funds from MCTES.