Graph-Based Entity-Oriented Search:

A Unified Framework in Information Retrieval

José Luís da Silva Devezas*
INESC TEC and FEUP InfoLab
jld@fe.up.pt

*Doctoral Program in Computer Science of the Universities of Minho, Aveiro, and Porto (MAP-i)

The 42nd European Conference on Information Retrieval (ECIR 2020), April 14-17, Universidade NOVA de Lisboa, Lisbon, Portugal

Universidade do Minho

universidade
de aveiro

U.PORTO

Who am I?

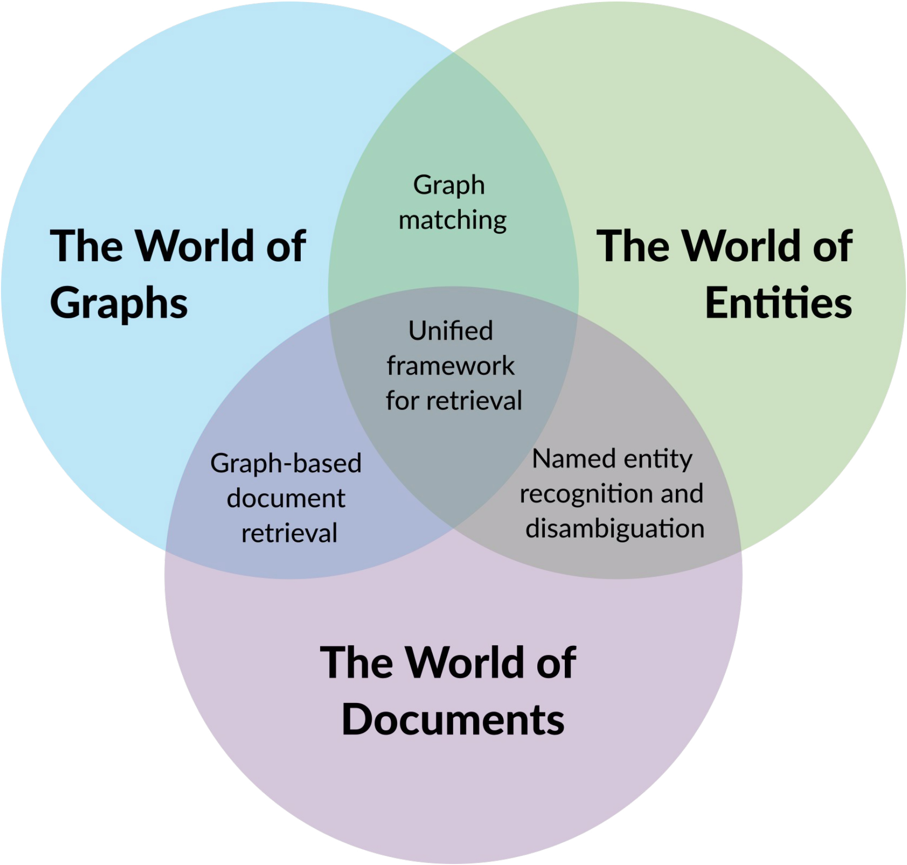
- 4th year PhD student at MAP-i, the Doctoral Program in Computer Science of the Universities of Minho, Aveiro, and Porto.
- Over the years, I have used graphs for diverse applications:

Link analysis for blog retrieval using h-index	Master's Thesis
Analysis of multidimensional entity co-occurrence networks from news clips	Breadcrumbs
Hybrid music recommendation	Juggle
Entity-oriented search engine for the University of Porto	ANT

- Currently, I am working with hypergraphs to build a general retrieval model, applied to entity-oriented search.

Unified framework for retrieval

- Important relations might only emerge from cross-referencing information at a low level.
- Heterogeneous data must be mapped to a common representation model.
- General retrieval should be able to solve any information need, over the represented data, through a universal ranking function.


The World of Graphs
Graph matching
The World of Entities
Unified framework for retrieval
Graph-based document retrieval
Named entity recognition and disambiguation
The World of Documents

Query: entertainers that are friends with astronauts who walked on the moon

"After the act, Kevin Foster went down to the audience, to hug his friend, Neil Armstrong, who had been sitting in the crowd since the beginning of the show."

Neil Armstrong	:isA	Astronaut
Neil Armstrong	:walkedOn	Moon
Buzz Aldrin	:isA	Astronaut
Buzz Aldrin	:walkedOn	Moon
Kevin Foster	:isA	Entertainer

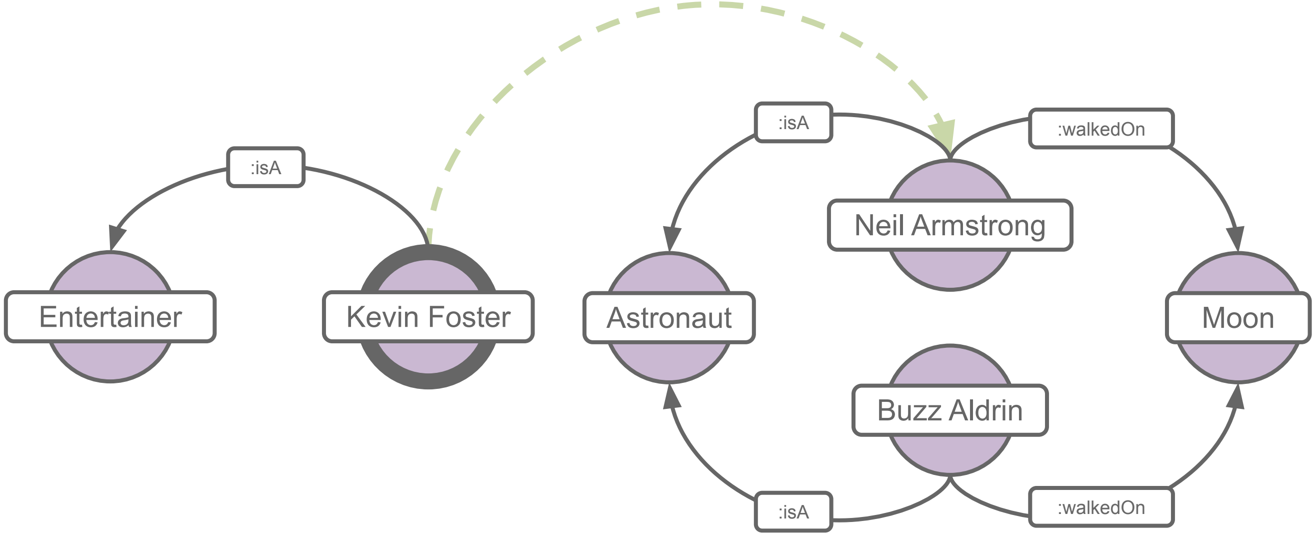


Illustration supported on the example by Hannah Bast and Björn Buchhold on "An Index for Efficient Semantic Full-text Search" (2013):

«Consider the query for entertainers that are friends with [an astronaut who walked on the moon] and that the fact about friendship is retrieved from the text and not part of our ontology. There is no way to process the full-text and ontology part independently and afterwards combine the results.»

Graphs as the common denominator

- Text as a graph (e.g., graph-of-word).
- Knowledge as a graph (e.g., RDF graph).
- Image or audio as a similarity graph.
- Graphs as the data structure for a general representation model.

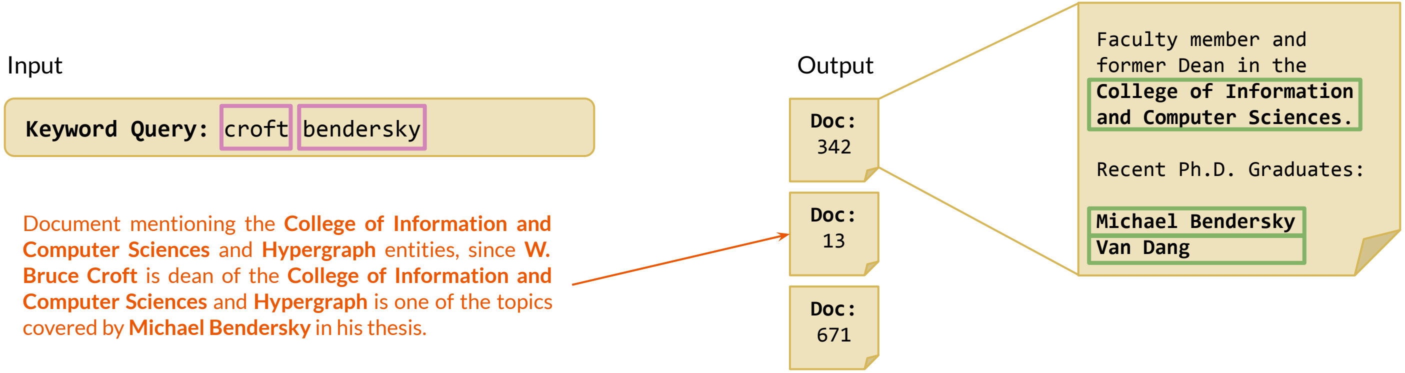
Graph-based entity-oriented search

- Application to entity-oriented search,
- Leading towards general information retrieval,
- Exploring graph-based models as a solution.

> *Entity-oriented search is the search paradigm of organizing and accessing information centered around entities, and their attributes and relationships.*
>
> – Krisztian Balog, Entity-Oriented Search, 2018.

Four retrieval tasks

Ad hoc document retrieval (leveraging entities)



Input

Keyword Query: croft bendersky

Document mentioning the **College of Information and Computer Sciences** and **Hypergraph** entities, since **W. Bruce Croft** is dean of the **College of Information and Computer Sciences** and **Hypergraph** is one of the topics covered by **Michael Bendersky** in his thesis.

Output

Doc: 342

Doc: 13

Doc: 671

Faculty member and former Dean in the **College of Information and Computer Sciences.**

Recent Ph.D. Graduates:

Michael Bendersky
Van Dang

Ad hoc entity retrieval

Input

Keyword Query: croft bendersky

Output

Entity: [Person] W. Bruce Croft

Entity: [Person] Michael Bendersky

Related entity finding

Input

Entity: `[Person]` Michael Bendersky

Type: `[ScholarlyArticle]`

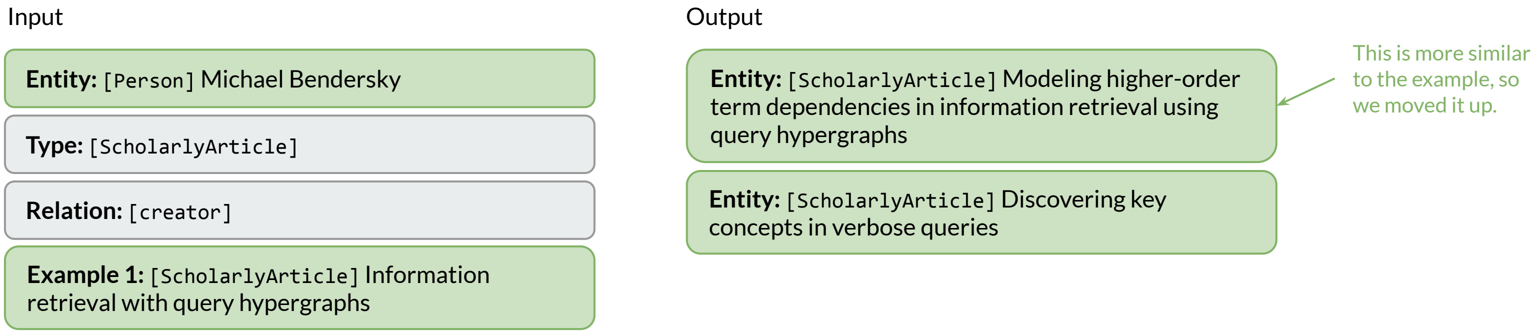
Relation: `[creator]`

Output

Entity: `[ScholarlyArticle]` Discovering key concepts in verbose queries

Entity: `[ScholarlyArticle]` Modeling higher-order term dependencies in information retrieval using query hypergraphs

Entity list completion



Input

Entity: `[Person]` Michael Bendersky

Type: `[ScholarlyArticle]`

Relation: `[creator]`

Example 1: `[ScholarlyArticle]` Information retrieval with query hypergraphs

Output

Entity: `[ScholarlyArticle]` Modeling higher-order term dependencies in information retrieval using query hypergraphs

Entity: `[ScholarlyArticle]` Discovering key concepts in verbose queries

This is more similar to the example, so we moved it up.

Thesis Statement

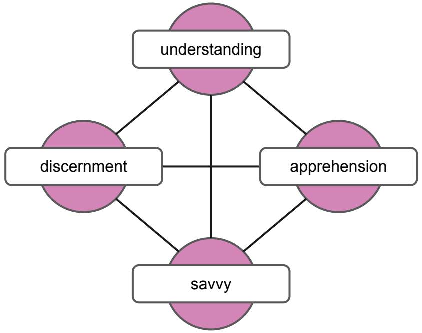
A graph-based joint representation of unstructured and structured data has the potential to unlock novel ranking strategies, that are, in turn, able to support the generalization of entity-oriented search tasks and to improve overall retrieval effectiveness by incorporating explicit and implicit information derived from the relations between text found in corpora and entities found in knowledge bases.

– José Devezas, Graph-Based Entity-Oriented Search, 2020.

We talk about graphs, but...

...we quickly switched to hypergraphs.

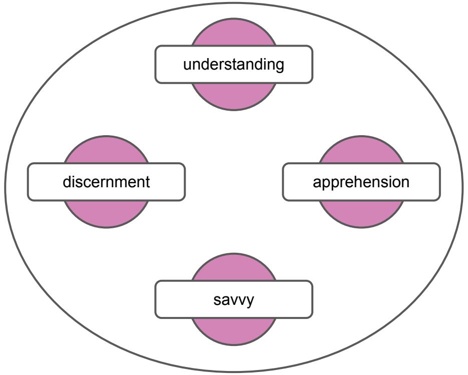
- Graphs did not scale well
- Hypergraphs are more expressive
- And it's easier to limit complexity during model design.



Synonymy modeled as an undirected graph.

BFS: O(|V| + |E|) = O(b^{d})

b ⇒ branching factor (outdegree)
d ⇒ distance (diameter)



Synonymy modeled as an undirected hypergraph.

O(|V|)

Complexity is linear in both cases, however the number of nodes not only converges (there is only a limited number of terms and entities), but also, when compared to the number of edges, it is sqrt(|E|) at worst.

GoW

Document-based Binary relations

Captures context by linking each term to its following terms within a sliding window of size n.

- INEX 2009 Wikipedia Collection (subset based on 10 topics)
  - 7,487 documents (29.1x)
  - 492,185 vertices (61.9x)
  - 22,906,803 edges (312.8x)
  - |E| = 46.5 x |V|

GoE

Collection-based Binary relations

Captures term sequences, term-entity relations based on substring matching and entity-entity relations based on a set of triples.

- INEX 2009 Wikipedia Collection (subset based on 10 topics)
  - 7,487 documents (29.1x)
  - 981,647 vertices (67.5x)
  - 9,942,647 edges (202.5x)
  - |E| = 10.1 x |V|

HGoE

Collection-based n-ary relations

Captures set-relations: documents are sets of terms and entities; related entities, within a given context, are sets of entities; a set of terms are linked to its entities.

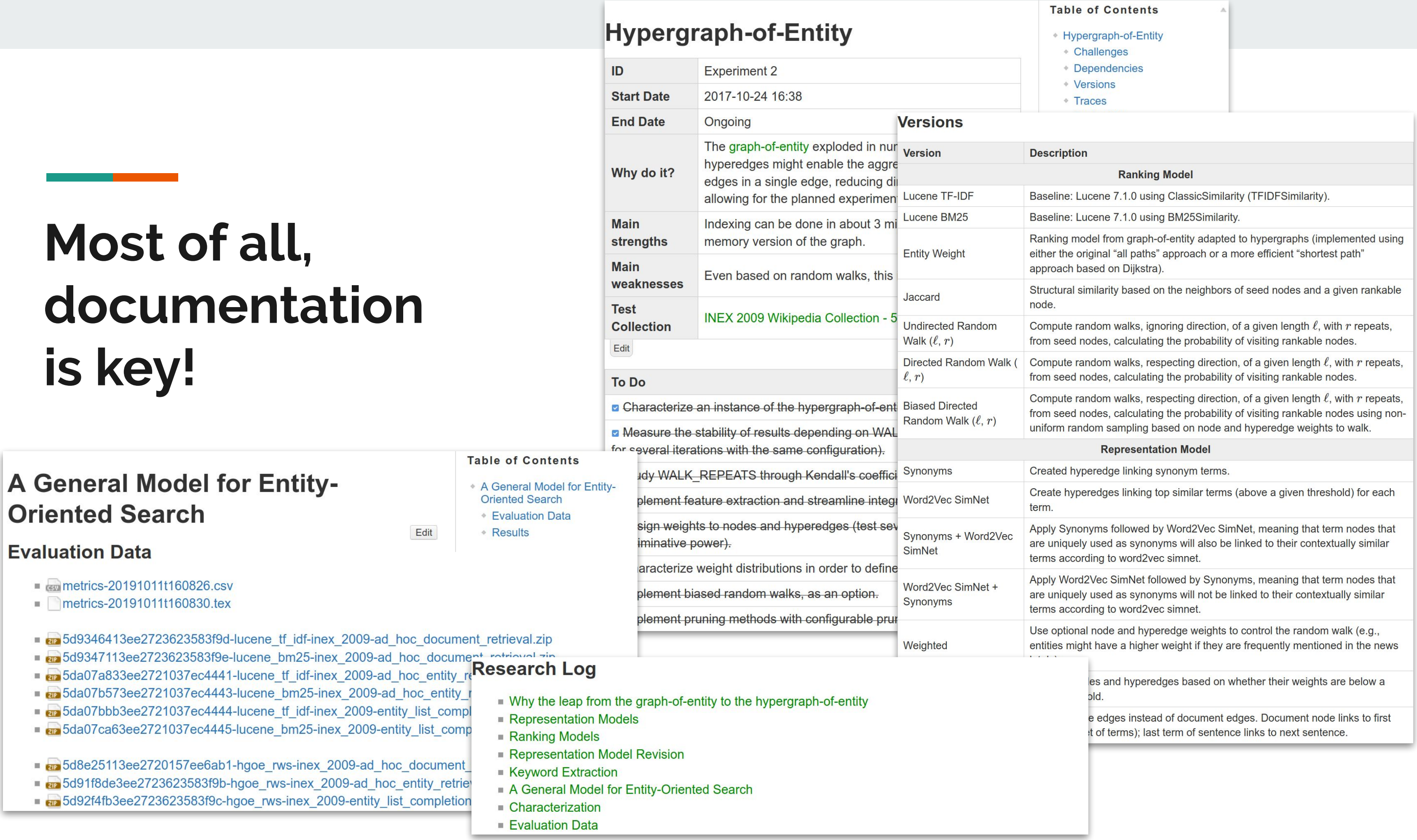
- INEX 2009 Wikipedia Collection (subset based on 10 topics)
  - 7,487 documents (29.1x)
  - 667,911 vertices (64.7x)
  - 295,921 hyperedges (86.5x)
  - |E| = 0.4 x |V|

Comparing the growth of graph-of-word (GoW), graph-of-entity (GoE) and hypergraph-of-entity (HGoE).

> *At the Balatonfüred Conference (1969), P. Erdös and A. Hajnal asked us why we would use hypergraphs for problems that can be also formulated in terms of graphs. The answer is that by using hypergraphs, one deals with generalizations of familiar concepts. Thus, hypergraphs can be used to simplify as well as to generalize.*
>
> – Claude Berge, Graphes et Hypergraphes, 1970.

Methodology and experiments

Most of all, documentation is key!



Hypergraph-of-Entity

ID	Experiment 2
Start Date	2017-10-24 16:38
End Date	Ongoing
Why do it?	The graph-of-entity exploded in nur hyperedges might enable the aggre edges in a single edge, reducing di allowing for the planned experimen
Main strengths	Indexing can be done in about 3 mi memory version of the graph.
Main weaknesses	Even based on random walks, this
Test Collection	INEX 2009 Wikipedia Collection - 5

Edit

To Do

- ☑ ~~Characterize an instance of the hypergraph-of-ent~~
- ☑ ~~Measure the stability of results depending on WAL for several iterations with the same configuration).~~
- ~~udy WALK_REPEATS through Kendall's coeffici~~
- ~~plement feature extraction and streamline integ~~
- ~~sign weights to nodes and hyperedges (test sev iminative power).~~
- aracterize weight distributions in order to define
- ~~plement biased random walks, as an option.~~
- ~~plement pruning methods with configurable prur~~

Versions

Version	Description
	Ranking Model
Lucene TF-IDF	Baseline: Lucene 7.1.0 using ClassicSimilarity (TFIDFSimilarity).
Lucene BM25	Baseline: Lucene 7.1.0 using BM25Similarity.
Entity Weight	Ranking model from graph-of-entity adapted to hypergraphs (implemented using either the original "all paths" approach or a more efficient "shortest path" approach based on Dijkstra).
Jaccard	Structural similarity based on the neighbors of seed nodes and a given rankable node.
Undirected Random Walk (ℓ, r)	Compute random walks, ignoring direction, of a given length ℓ, with r repeats, from seed nodes, calculating the probability of visiting rankable nodes.
Directed Random Walk ($\ell, r)$	Compute random walks, respecting direction, of a given length ℓ, with r repeats, from seed nodes, calculating the probability of visiting rankable nodes.
Biased Directed Random Walk (ℓ, r)	Compute random walks, respecting direction, of a given length ℓ, with r repeats, from seed nodes, calculating the probability of visiting rankable nodes using non-uniform random sampling based on node and hyperedge weights to walk.
	Representation Model
Synonyms	Created hyperedge linking synonym terms.
Word2Vec SimNet	Create hyperedges linking top similar terms (above a given threshold) for each term.
Synonyms + Word2Vec SimNet	Apply Synonyms followed by Word2Vec SimNet, meaning that term nodes that are uniquely used as synonyms will also be linked to their contextually similar terms according to word2vec simnet.
Word2Vec SimNet + Synonyms	Apply Word2Vec SimNet followed by Synonyms, meaning that term nodes that are uniquely used as synonyms will not be linked to their contextually similar terms according to word2vec simnet.
Weighted	Use optional node and hyperedge weights to control the random walk (e.g., entities might have a higher weight if they are frequently mentioned in the news
	es and hyperedges based on whether their weights are below a old.
	e edges instead of document edges. Document node links to first t of terms); last term of sentence links to next sentence.

A General Model for Entity-Oriented Search

Edit

Evaluation Data

- metrics-20191011t160826.csv
- metrics-20191011t160830.tex

- 5d9346413ee2723623583f9d-lucene_tf_idf-inex_2009-ad_hoc_document_retrieval.zip
- 5d9347113ee2723623583f9e-lucene_bm25-inex_2009-ad_hoc_document_retrieval.zip
- 5da07a833ee2721037ec4441-lucene_tf_idf-inex_2009-ad_hoc_entity_re
- 5da07b573ee2721037ec4443-lucene_bm25-inex_2009-ad_hoc_entity_r
- 5da07bbb3ee2721037ec4444-lucene_tf_idf-inex_2009-entity_list_compl
- 5da07ca63ee2721037ec4445-lucene_bm25-inex_2009-entity_list_comp

- 5d8e25113ee2720157ee6ab1-hgoe_rws-inex_2009-ad_hoc_document_
- 5d91f8de3ee2723623583f9b-hgoe_rws-inex_2009-ad_hoc_entity_retriev
- 5d92f4fb3ee2723623583f9c-hgoe_rws-inex_2009-entity_list_completion

Research Log

- Why the leap from the graph-of-entity to the hypergraph-of-entity
- Representation Models
- Ranking Models
- Representation Model Revision
- Keyword Extraction
- A General Model for Entity-Oriented Search
- Characterization
- Evaluation Data

Five stages of experimentation

1. **Conception**
   - Where we tested graph-based models (inspired by the graph-of-word, we proposed the graph-of-entity);
   - But this didn’t scale even to support our experiments, much less for a production search engine.
2. **Representation model**
   - We proposed the hypergraph-of-entity and experimented with multiple variations:
     - Text-only;
     - Using internal and external knowledge (i.e., entities and relations);
     - Including synonyms from WordNet;
     - Including contextual similarity relations based on a word2vec similarity graph;
     - Introducing the concept of term frequency through TF-bins.

Five stages of experimentation

3. **Retrieval model**
    - We experimented with multiple parameterizations of the random walk score:
        - Length of the random walk;
        - Number of repeated random walks per seed node;
        - Number of cycles of node and edge fatigue;
        - Query expansion to neighboring nodes;
        - Considering or ignoring direction;
        - Biased or weighted random walks.
4. **Fallback to classical**
    - Tested the idea of fatigued random walks by proposing Fatigued PageRank;
    - Used classical link analysis metrics that were combined with a text-based score.

Five stages of experimentation

5. **Generalization testing**
   - Evaluated the performance of a unified retrieval framework over three entity-oriented search tasks:
     - Ad hoc document retrieval
     - Ad hoc entity retrieval
     - Entity list completion

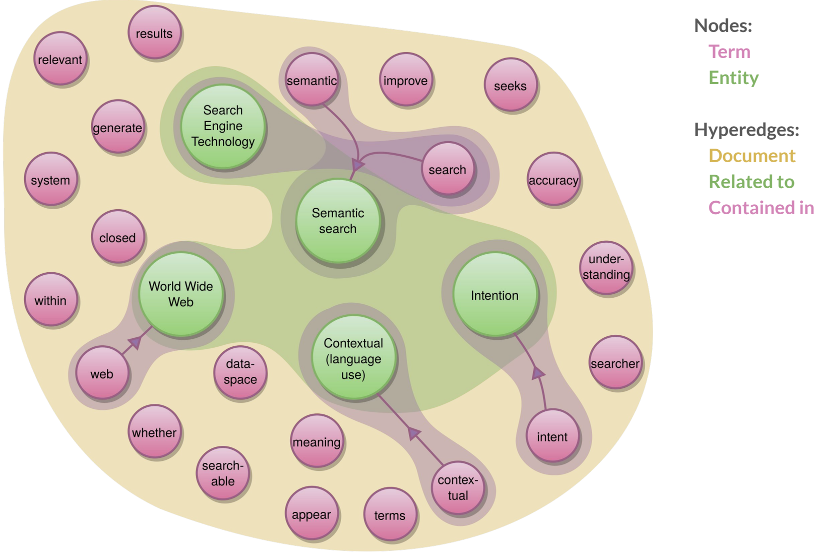
A general solution

Devezas, J., and S. Nunes (2019). Hypergraph-of-entity: A unified representation model for the retrieval of text and knowledge. In Open Computer Science (Topical Issue on Intelligent Methods for Textual Information Retrieval), 9(1), pp.103-127.

Hypergraph-of-entity

Representation model

- Collection-based hypergraph.
- Joint representation of terms, entities and their relations.
- For indexing combined data (e.g., corpora linked to knowledge bases).



Devezas, J., and S. Nunes (2019). Hypergraph-of-entity: A unified representation model for the retrieval of text and knowledge. In Open Computer Science (Topical Issue on Intelligent Methods for Textual Information Retrieval), 9(1), pp.103-127.

Hypergraph-of-entity

Representation model

We can easily add hyperedges for:

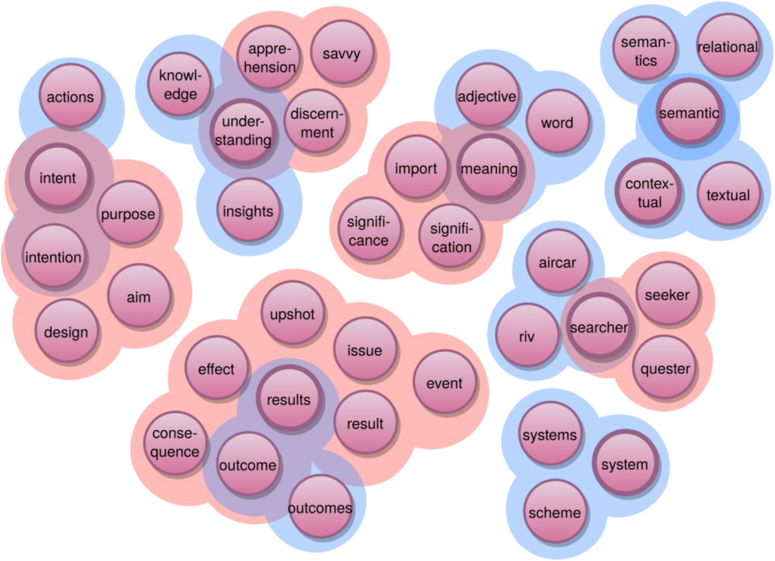
Synonyms

- Based on WordNet SynSets.

Contextual similarity

- Based on word2vec similarities.

Note: The remaining hyperedges and nodes are not displayed here to improve legibility.

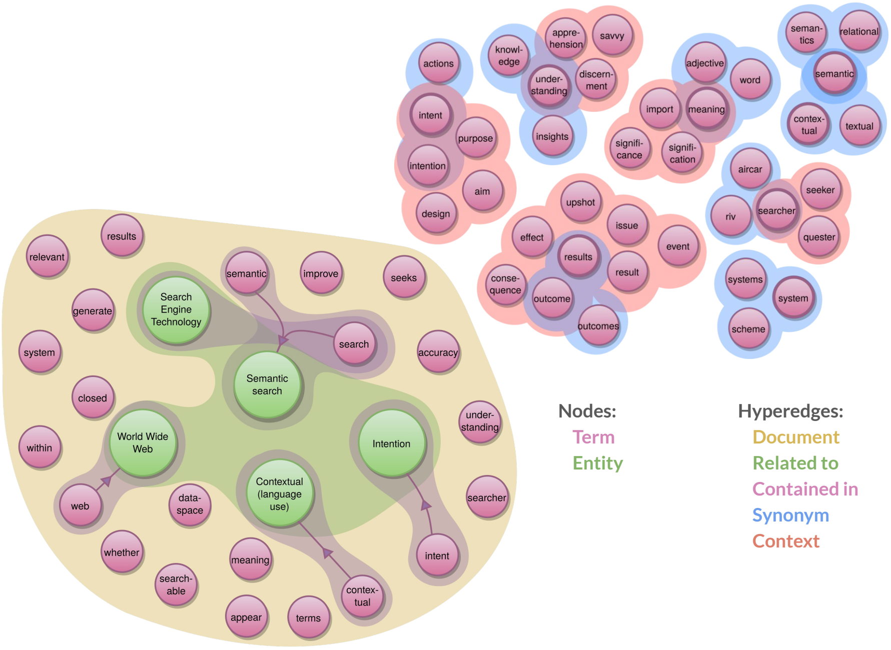


Devezas, J., and S. Nunes (2019). Hypergraph-of-entity: A unified representation model for the retrieval of text and knowledge. In Open Computer Science (Topical Issue on Intelligent Methods for Textual Information Retrieval), 9(1), pp.103-127.

Hypergraph-of-entity

Retrieval model

- Random walks departing from seeds, representing the query, assign a visitation frequency to nodes and hyperedges.
- Depending on the task, input and output changes, but the ranking function remains unchanged.
- Nodes or hyperedges are collected and ordered by visitation frequency to provide a ranking.



Discussion pointers

The model is inefficient for production...

- In fact, we could not even index the complete INEX 2009 Wikipedia collection without reducing documents to their top keywords.
- And queries were quite slow (~1-10m).
- Should we:
    - Attempt to prune the hypergraph? By which criteria?
    - Explore hypergraph embeddings? But this is a general mixed hypergraph… Can we do it?
    - Attempt to split the collection-based hypergraph, without losing information?

The model is not up to par in effectiveness...

- Yes, we are limited by inefficiency for further testing…
- But also, how could we ensure a timely indexing process based on information from:
  - Sentences;
  - Syntactic dependencies;
  - Multiple languages (i.e., for cross-language retrieval).
- That is, even the preprocessing pipeline can have a high impact in performance, when further expanding the model.
- A general problem for a general model?

But we should not give up...

- The goal of devising a unified framework for solving information needs is worth it!
- Overall the effectiveness for hypergraph-of-entity is lower than the state of the art.
- But it solves multiple tasks with a universal ranking function.
- Resulting in a balanced decrease in performance, when looking at each task's effectiveness compared to other solutions.
- I believe we have sufficient evidence to justify further pursuing this new idea!

Thank you!

You can learn more about the hypergraph-of-entity here:
https://doi.org/10.1515/comp-2019-0006
https://doi.org/10.1007/978-3-030-36683-4_1

Extra Slides

Hypergraph expressiveness

Strengths:

- n-ary relations.
- Hierarchical relations.
- Overlapping relations.
- Weighted nodes and hyperedges.
- Directed dependencies between two groups.

Weaknesses:

- No membership degree (i.e., a node either belongs to a hyperedge or it doesn’t).
- Non-explicit dependencies between groups.

Five stages of experimentation

5. **Generalization testing**
    - We relied on three distinct Lucene baselines, supported on:
        - Two representation models:
            - An index based on text from documents;
            - An index based on entity profiles built from sentences mentioning the entity.
        - Two retrieval approaches:
            - Regular keyword query (TF-IDF / BM25);
            - Using `MoreLikeThis` for inter-document similarity.

Devezas, J., and S. Nunes (2019). Hypergraph-of-entity: A unified representation model for the retrieval of text and knowledge. In Open Computer Science (Topical Issue on Intelligent Methods for Textual Information Retrieval), 9(1), pp.103-127.

Hypergraph-of-entity

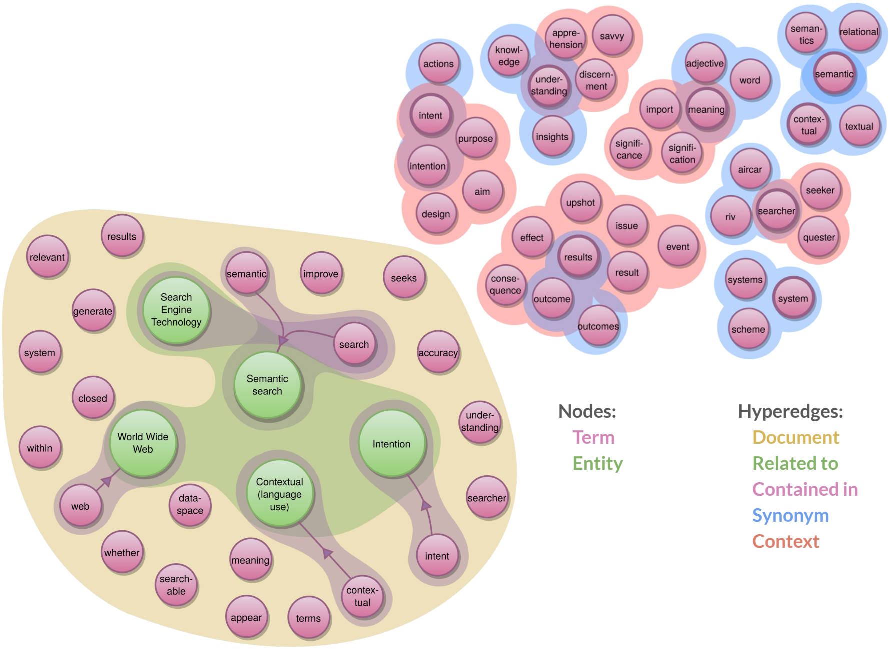
- The hypergraph-of-entity is:
  - A general model for entity-oriented search;
  - A joint representation model for corpora and knowledge bases.
- Its random walk score is a universal ranking function for:
  - Ad hoc document retrieval;
  - Ad hoc entity retrieval;
  - Related entity finding;
  - Entity list completion.

Devezas, J., and S. Nunes (2019). Hypergraph-of-entity: A unified representation model for the retrieval of text and knowledge. In Open Computer Science (Topical Issue on Intelligent Methods for Textual Information Retrieval), 9(1), pp.103-127.

Hypergraph-of-entity

Retrieval model

- Supports the generalization of tasks, through the random walk score.
- Query term nodes can be used as seeds (either directly or expanded to entities).
- We can query by terms or entities...
- ...And rank documents, entities, or even terms.



results
relevant
generate
system
closed
within
web
whether
searchable
appear
terms
meaning
contextual
intent
searcher
understanding
accuracy
seeks
improve
semantic
search
Search Engine Technology
Semantic search
World Wide Web
data-space
Contextual (language use)
Intention
actions
intent
purpose
intention
aim
design
knowledge
apprehension
savvy
understanding
discernment
insights
adjective
word
import
meaning
significance
signification
semantics
relational
semantic
contextual
textual
aircar
riv
searcher
seeker
quester
upshot
issue
effect
event
results
result
consequence
outcome
outcomes
systems
system
scheme
Nodes:
Term
Entity
Hyperedges:
Document
Related to
Contained in
Synonym
Context