Mapping Terms and Entities with Hypergraphs

Creative Collab 2019   **Last update:** June 18, 2019

José Devezas   MAP-i 2016/2017   INESC TEC & FEUP InfoLab

Supervisor: Sérgio Nunes   U.Porto
**External advisor:**   Bruno Martins   U.Lisboa

**Universidade do Minho**

universidade de aveiro

U.PORTO

Hypergraph

- A graph models dyadic relations between two elements.

- A hypergraph is a generalization of a graph, where edges (or rather hyperedges) are able to establish n-ary relations between an arbitrary number of elements.

- Hyperedges can be:

  - Undirected (a set of nodes) $\{n_1, n_2, n_3\}$

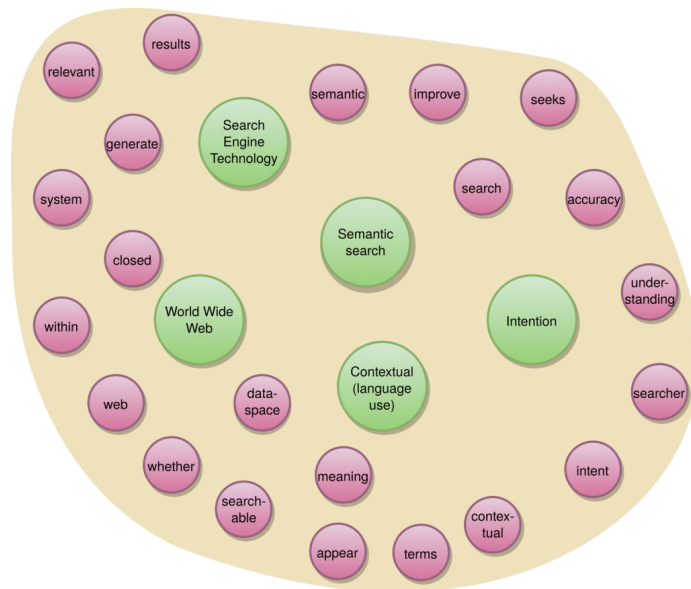 - Or directed (a pair of sets of nodes) $(\{n_1, n_2\}, \{n_3\})$
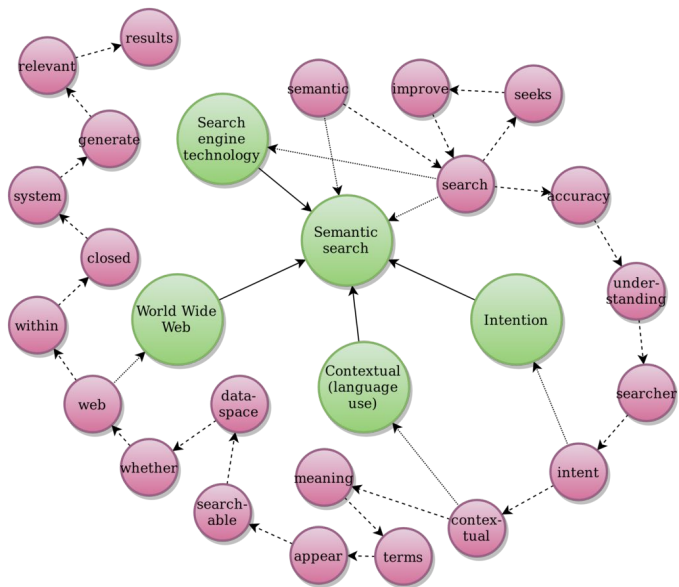
Graph: dyadic/binary relations.
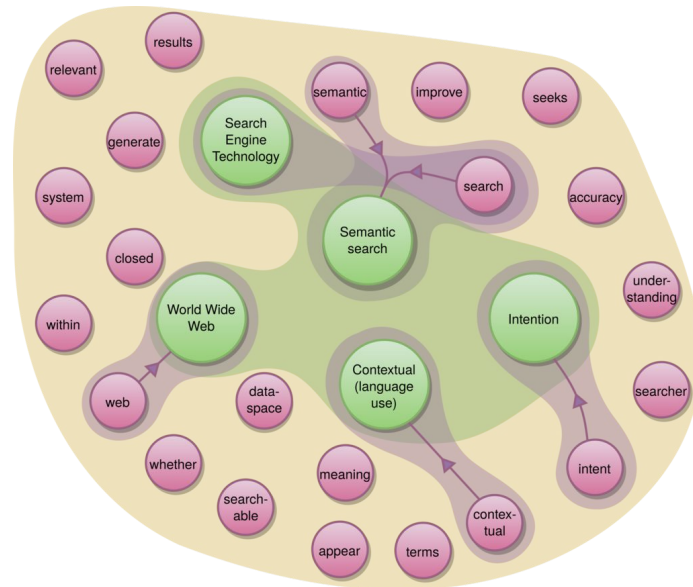
Hypergraph: n-ary relations.
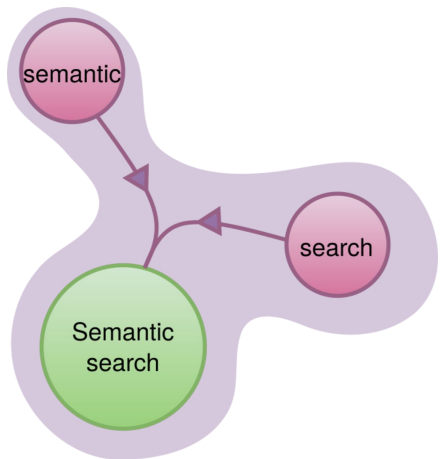
Directed hyperedges
(e.g., contained in).

Undirected hyperedges
(e.g., a bag of words and entities).

What can we represent with hypergraphs? Are there limitations?

Hypergraph expressiveness

Strengths:

- n-ary relations.

- Hierarchical relations.

- Overlapping relations.

- Weighted nodes and hyperedges.
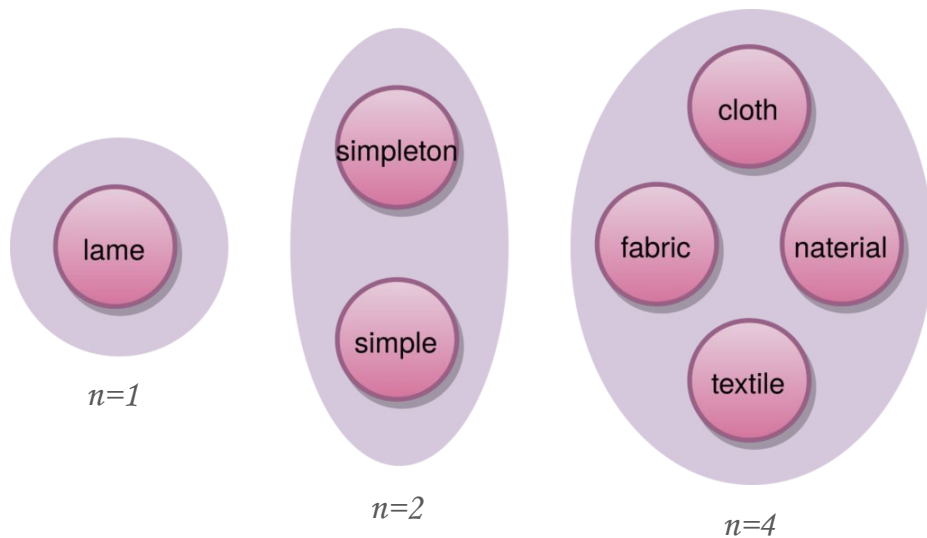
- Directed dependencies between two groups.

Weaknesses:

- No membership degree (i.e., a node either belongs to a hyperedge or it doesn't).

- Non-explicit dependencies between groups.
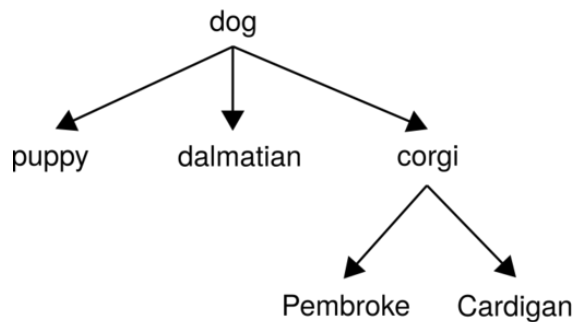
Strengths

n-ary relations

$n=1$

$n=2$

$n=4$

Hyperedge cardinality: Number of nodes n that belong to a hyperedge.

Hierarchical relations
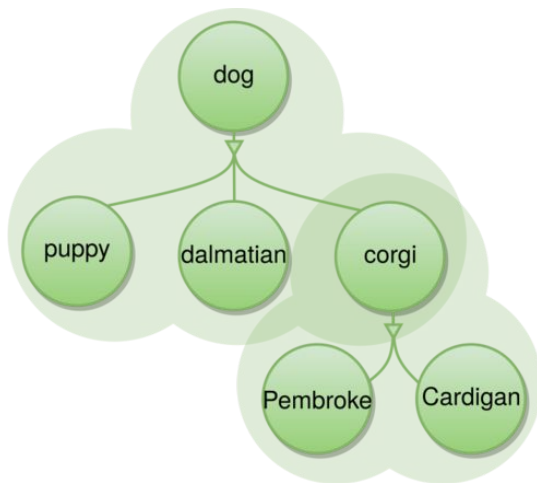
e.g., hyponyms: Two levels of hyponymy, starting from 'dog', and then from 'corgi'.
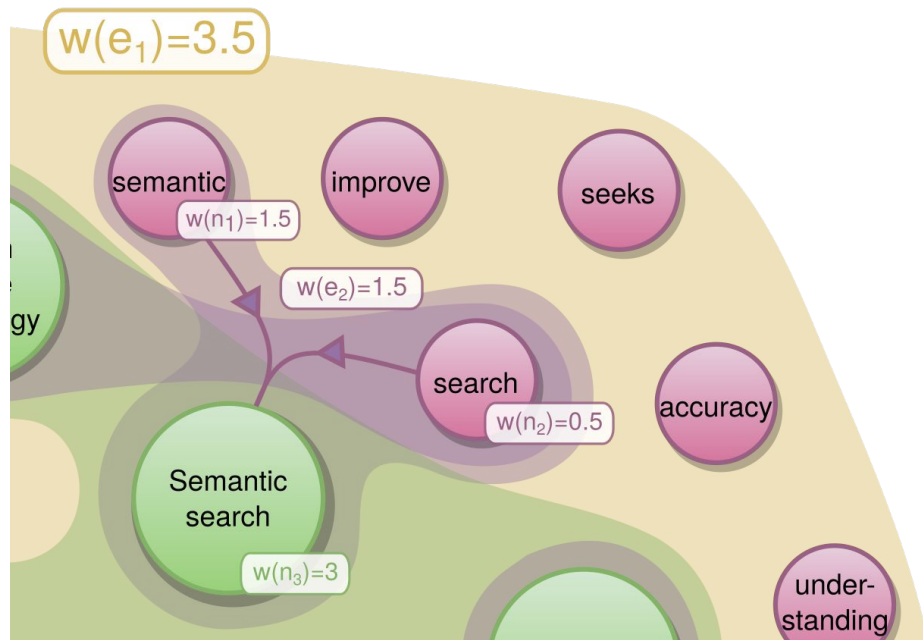
Overlapping relations
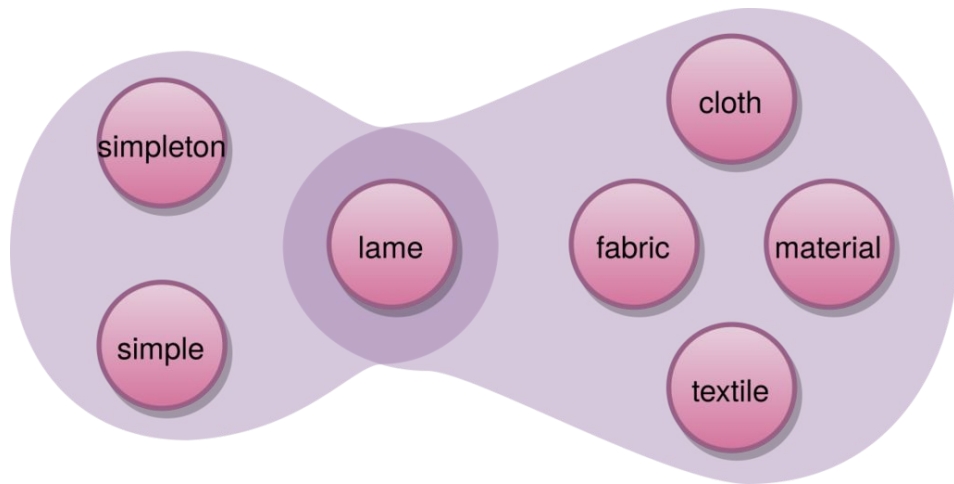
e.g., synonyms: 'lame' has two senses that can be represented by two overlapping undirected hyperedges.

Weighted nodes and hyperedges

Weighted hypergraph: different degrees of importance can be assigned to nodes and hyperedges. These are globally set for the hypergraph.

Directed dependencies between two groups

- While we still have an n-ary relation…

- …The dependency is binary.

- It's between a group of two terms ('semantic' and 'search')…

- …And a group of one entity ('Semantic search').
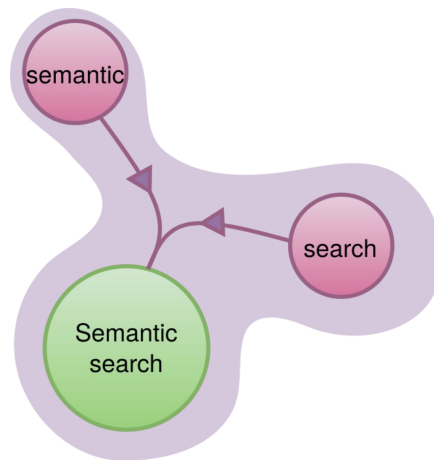
$$(\{semantic, search\}, \{Semantic\ search\})$$

Weaknesses

...and how to mitigate them.

No membership degree
...solved through fuzzy hypergraphs.

Undirected hyperedges:

$$\{lame, simpleton, simple\}$$

$$\{lame, fabric, cloth, material, textile\}$$
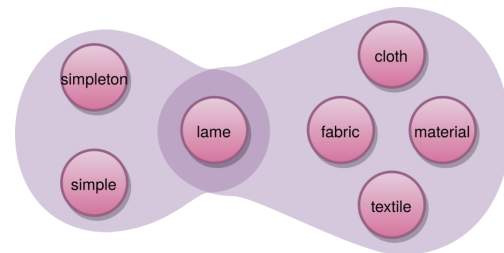
- A node either belongs to a hyperedge or it doesn't.

- In order to set a membership degree, we need a fuzzy hypergraph (Lee-Kwang and Lee, 1995).

No membership degree
...solved through fuzzy hypergraphs.
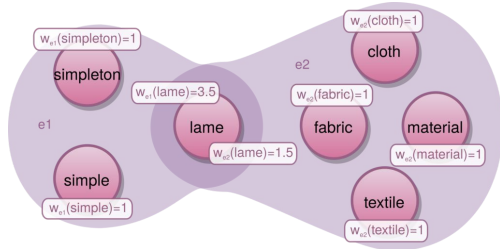
Undirected fuzzy hyperedges:

$$\{(lame, w(lame) = 3.5),$$
$$(simpleton, w(simpleton) = 1),$$
$$(simple, w(simple) = 1)\}$$

$$\{(lame, w(lame) = 1.5),$$
$$(fabric, w(fabric) = 1),$$
$$(cloth, w(cloth) = 1),$$
$$(material, w(material) = 1),$$
$$(textile, w(textile) = 1)\}$$

- Where instead of defining a hyperedge as a set of nodes...

- ...We define a fuzzy hyperedge as a set of pairs of node and weight function.

Non-explicit dependencies between groups
...mitigated through hierarchical combinations of hyperedges.

- Directed hyperedges are able to establish dependencies between two groups of nodes.

- But how do we model the dependencies between $n > 2$ groups?
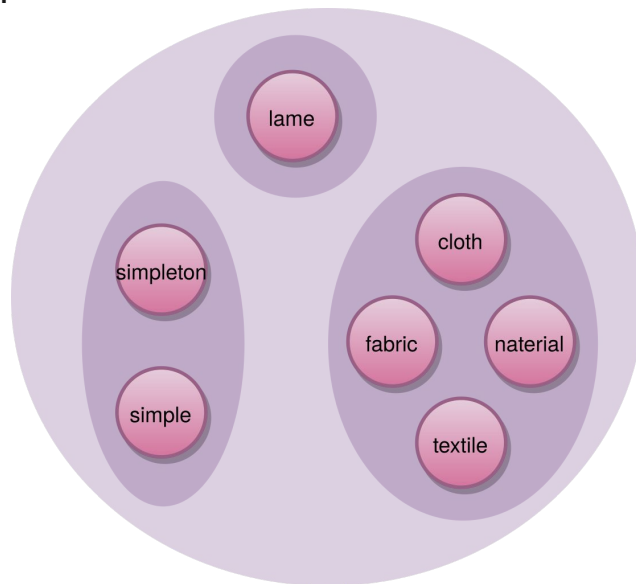
Non-explicit dependencies between groups

...mitigated through hierarchical combinations of hyperedges.

- One way would be to define a hyperedge that contains the n hyperedges for the groups.

- Although a hyperedge cannot directly contain other hyperedges, it can contain their nodes.

- Subsumption is then verifiable rather than explicitly represented.
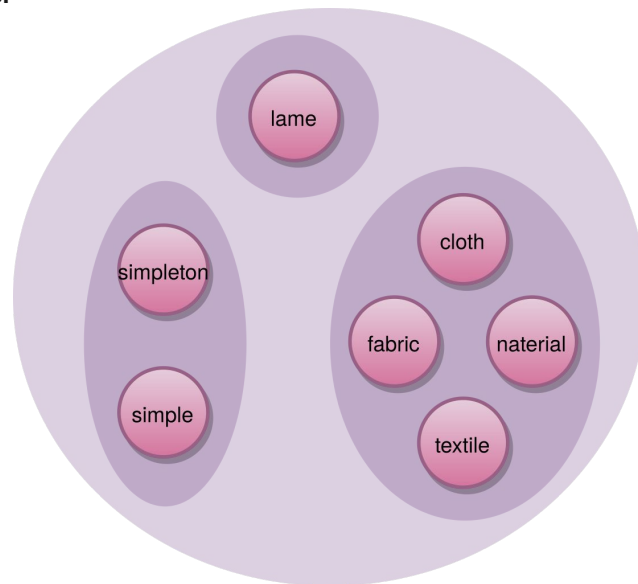
Has anyone used hypergraphs for the representation or retrieval of documents?

Garshol, L. M. (2004). Metadata? Thesauri? Taxonomies? Topic maps! Making sense of it all. Journal of Information Science, 30(4), 378-391.

Topics maps

- One of the first hypergraph-based models for indexing documents.

- Built to merge back-of-the-book indices.

- And to improve findability.

- In practice, though, few work exists on topics maps applied to search.

Nodes:

- **Topics** (e.g., entities, types);
- **Occurrences** (information resources, e.g., document identifier+page).

Edges:

- Binary connections between topics and occurrences;
- **Associations**: n-ary connections between topics.

Query hypergraph

Bendersky, M., & Croft, W. B. (2012, August). Modeling higher-order term dependencies in information retrieval using query hypergraphs. In Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval (pp. 941-950). ACM.

Nodes:

- **Concept** (terms, bigrams, noun phrases, named entities, dependencies);
- **Document**.

Hyperedges:

- **Local** (between each concept and the document).
- **Global** (between all concepts and a document).

- Ranking based on the **factor graph** representation of the hypergraph.

- Score computed through:

  - The combination of **log-linear models** for local and global factors;

  - Based on a **language modeling estimate** for each concept.

Dekker, R. H., & Birnbaum, D. J. (2017). It's more than just overlap: Text As Graph. In Proceedings of Balisage: The Markup Conference (Vol. 19).

TAG: Text As Graph

- Proposed by Dekker and Birnbaum in 2017 as a hypergraph-based representation model for text.

- Built to mitigate XML limitations (e.g., lack of overlap).

Nodes:

- Document;
- Text;
- Markup;
- Annotation.

Hyperedges:

- Document-to-text;
  - Marks the beginning of the document.

- Text-to-text;
  - Links sequences of tokens.

- Annotation-to-text/-markup /-annotation;
  - Links text blocks to parts of the document.

- Markup-to-text/-markup;
  - Assigns types (possibly hierarchical).

We also proposed a hypergraph-based model for the joint representation of text and entities, and generalized entity-oriented search.

What is entity-oriented search?

"Entity-oriented search is the search paradigm of organizing and accessing information centered around entities, and their attributes and relationships."

– Balog, Entity-Oriented Search, 2018.

Like Google!

Or Bing!

Or even ANT!

(Shameless advertising)

The tasks

Entity-oriented search tasks

Ad hoc document retrieval (using 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

Entity-oriented search tasks

Ad hoc entity retrieval

Input

| Keyword Query: | croft | bendersky |

Output

Entity: [Person] W. Bruce Croft

Entity: [Person] Michael Bendersky

Entity-oriented search tasks

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-oriented search tasks

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.

A general solution

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).
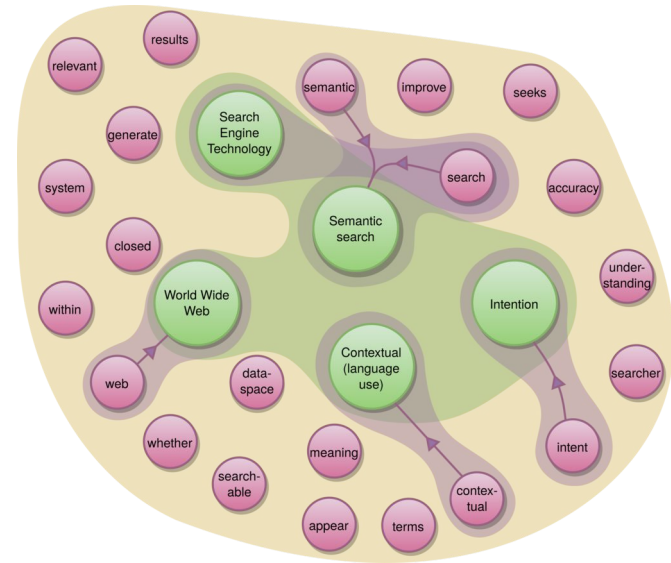
Hypergraph-of-entity

Representation model

Nodes:

- Terms;
- Entities.

Hyperedges:

- Document;
- Related to;
- Contained in.
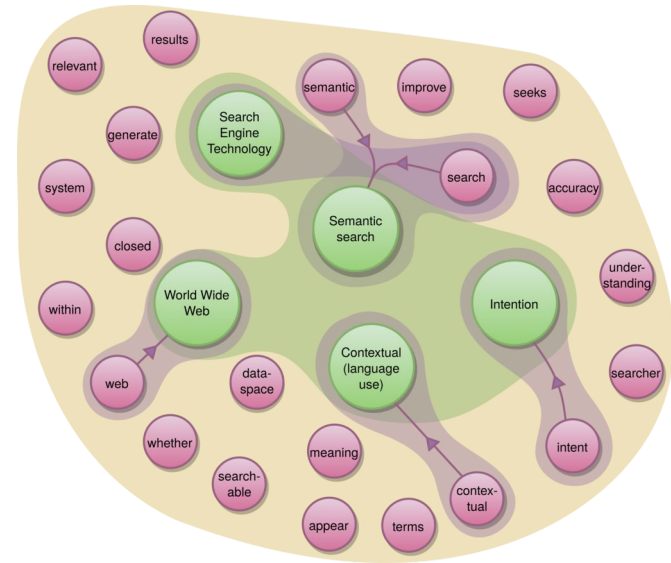
Hypergraph-of-entity

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.

Hypergraph-of-entity

Ranking 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 term or entities...

- ...And rank documents, entities, or even terms.
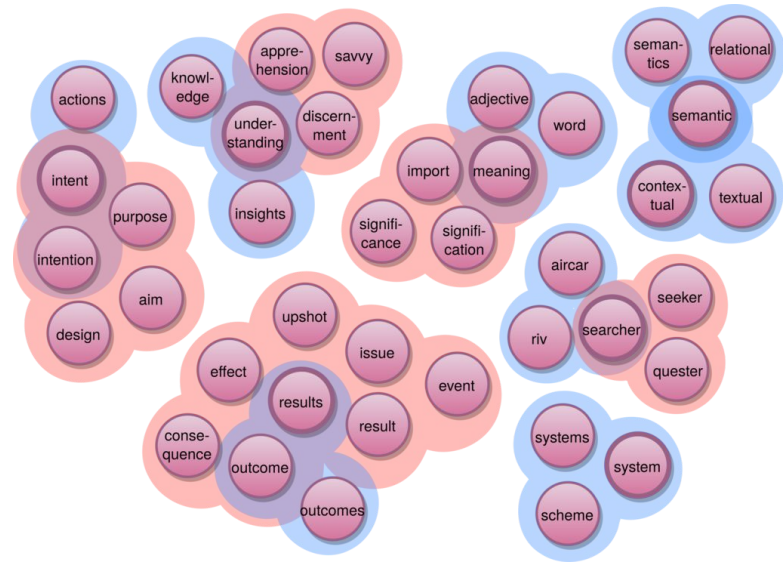
Hypergraph-of-entity

Ranking model

- Random walks departing from seeds 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.

Wrapping up...

Final remarks

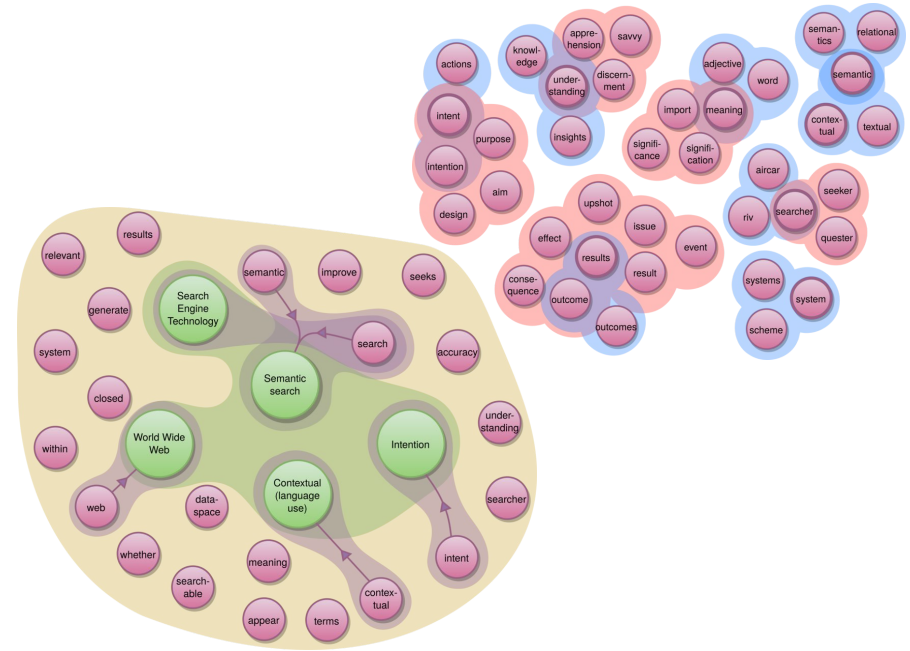- Hypergraphs are expressive and flexible data structures, ideal for the joint representation of unstructured corpora and structured knowledge bases.

- They are also a way to harness higher-order relations in information retrieval.

Final remarks

- In the hypergraph-of-entity, a document hyperedge is like a bag of words and entities. Adding synonym or contextual similarity hyperedges creates new semantic bridges between documents.

- Random walks in graphs have historical relevance in web search engines (e.g., PageRank). Applying them to hypergraphs was a natural step.

Thank you!

You can experiment with the hypergraph-of-entity and other retrieval models using our evaluation framework, Army ANT:

https://github.com/feup-infolab/army-ant

Also available as a Docker image:

https://github.com/feup-infolab/army-ant-install