Graph-of-Entity

A Model for Combined Data Representation and Retrieval

José Devezas*, Carla Lopes and Sérgio Nunes
INESC TEC and FEUP InfoLab
{jld, ctl, ssn}@fe.up.pt

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

8th Symposium on Languages, Applications and Technologies, SLATE 2019, June 27-28, Department of Mathematics, University of Coimbra, Portugal

Universidade do Minho

universidade de aveiro

U.PORTO

Information retrieval (IR) is finding material (usually documents) of an unstructured nature (usually text) that satisfies an information need from within large collections (usually stored on computers).

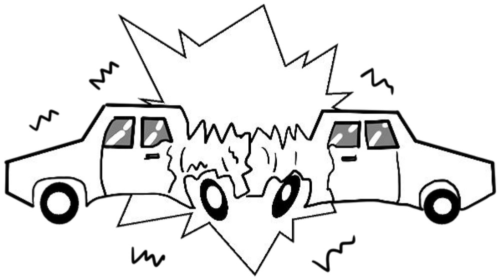
– Manning et al., Introduction to Information Retrieval, 2008.

10 years later....

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

– Balog, Entity-Oriented Search, 2018.

Two definitions collide



- **Classical information retrieval:**
  - Unstructured data;
  - Inverted index;
  - Partial structure through fields (e.g., for title, headers, etc.).
- **Entity-oriented search:**
  - Structured data;
  - Triplestore;
  - Partial full-text search (e.g., over objects of triples with a given predicate or graph).

How do we bridge the two concepts?

Let us look at combined data.

[...] <u>combined data</u> is obtained by one or both of the following two principles:

link: *link a text to a knowledge base by recognizing mentions of entities from the knowledge base in the text and linking to them*

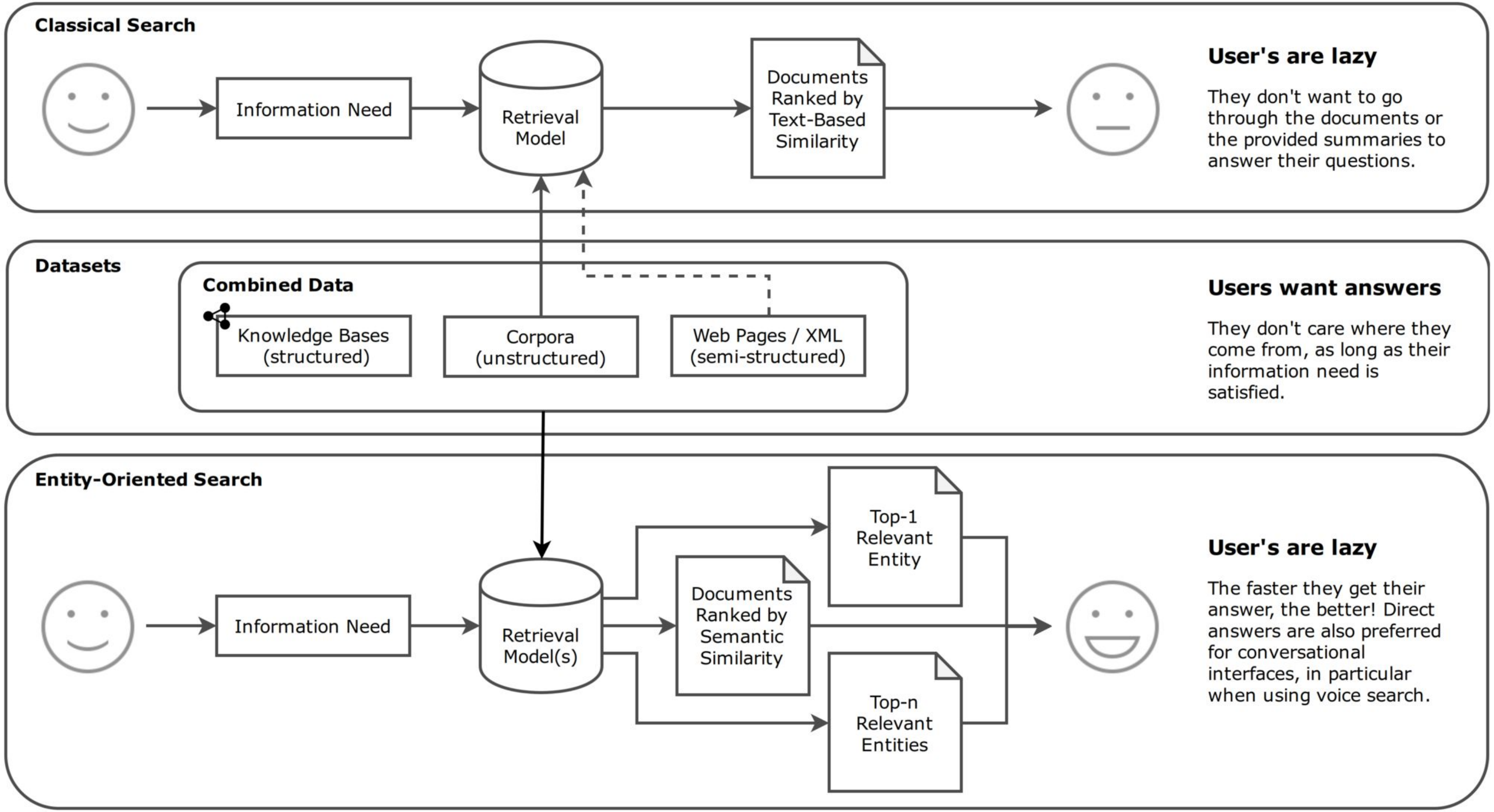
TXT ↔ KB

mult: *combine multiple knowledge bases with different naming schemes (such that the same entity or relation may exist with different names)*

KB ↔ KB

– Bast et al., Semantic Search on Text and Knowledge Bases, 2016.

In summary…

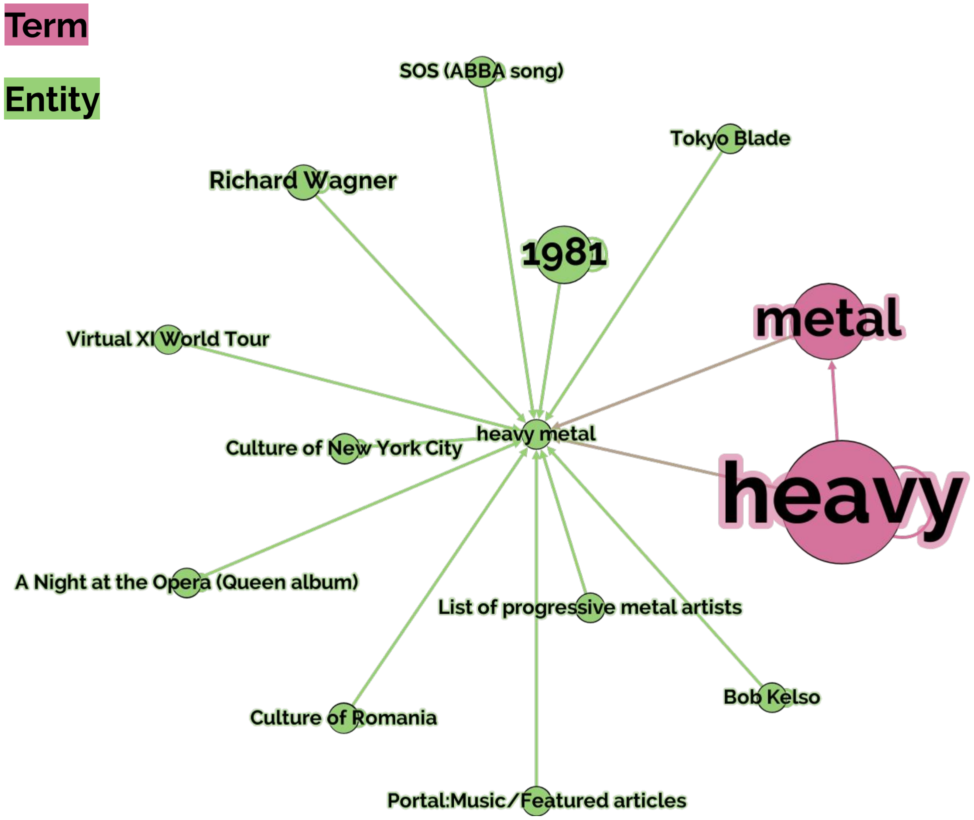


What information systems have in common is that they focus on the user. In order to provide the best solution to a user's information need, we should not only provide results from different information sources, but also be able to cross-reference that information.

How do we build a retrieval model for combined data?

Graphs!

- Knowledge bases are inherently graphs.
- But how can we represent text as a graph?
- And how do we combine text and knowledge bases as a graph?



Let us look into the literature.

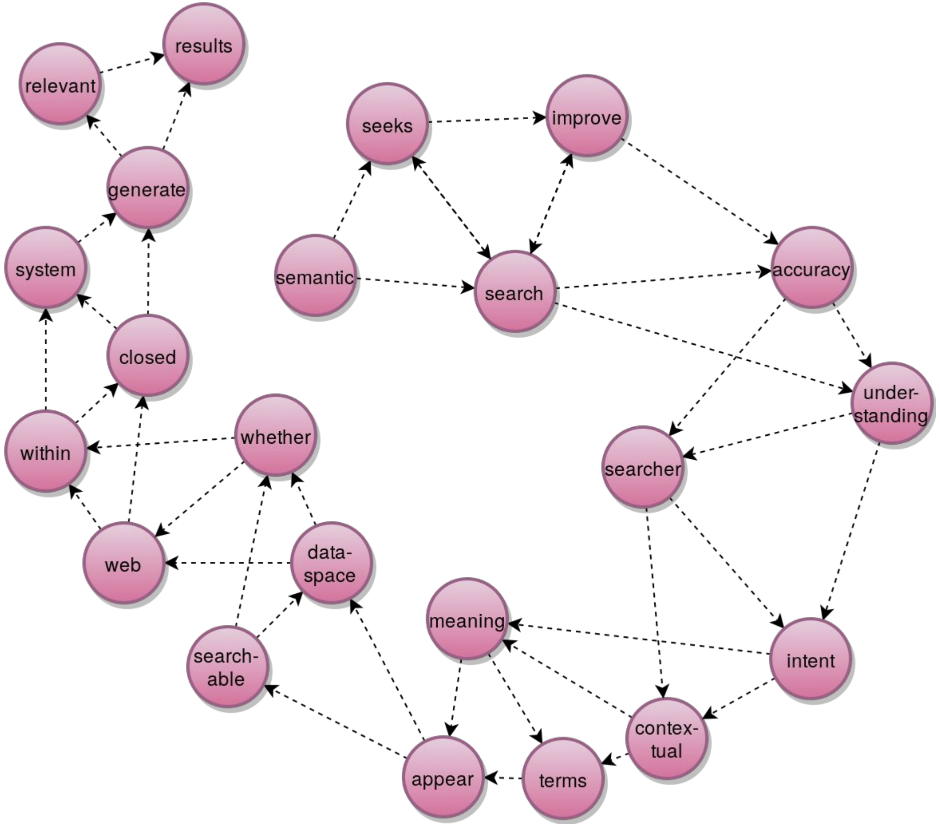
Graph-of-Word

Rousseau and Vazirgiannis (2013)

Graph-of-Word

Representation

- Document-based graph.
- Each term links to the following n terms.
- This establishes a context for each term.
- The graph can be discarded after computing the statistics.



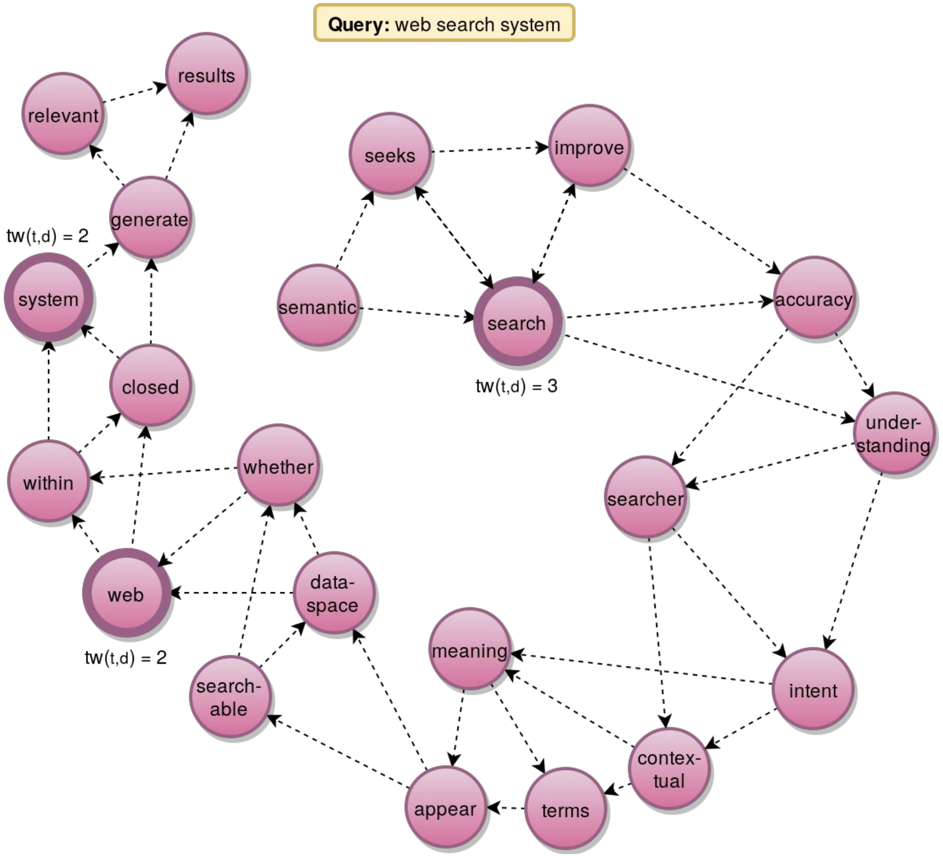
Graph-of-Word

Ranking

- Ranking is computed based on *TW-IDF*:

$$TW\text{-}IDF(t,d) = \frac{tw(t,d)}{1-b+b\times\frac{|d|}{avdl}} \times log\frac{N+1}{df(t)}$$

- $tw(t, d)$ is the indegree of term t in the graph-of-word for document d.
- This is divided by a pivoted document length normalization factor with $b = 0.003$.
- And multiplied by the IDF.



Ok, so text can be represented as a graph of “word contexts”!

Now let's try something similar, but also include entities.

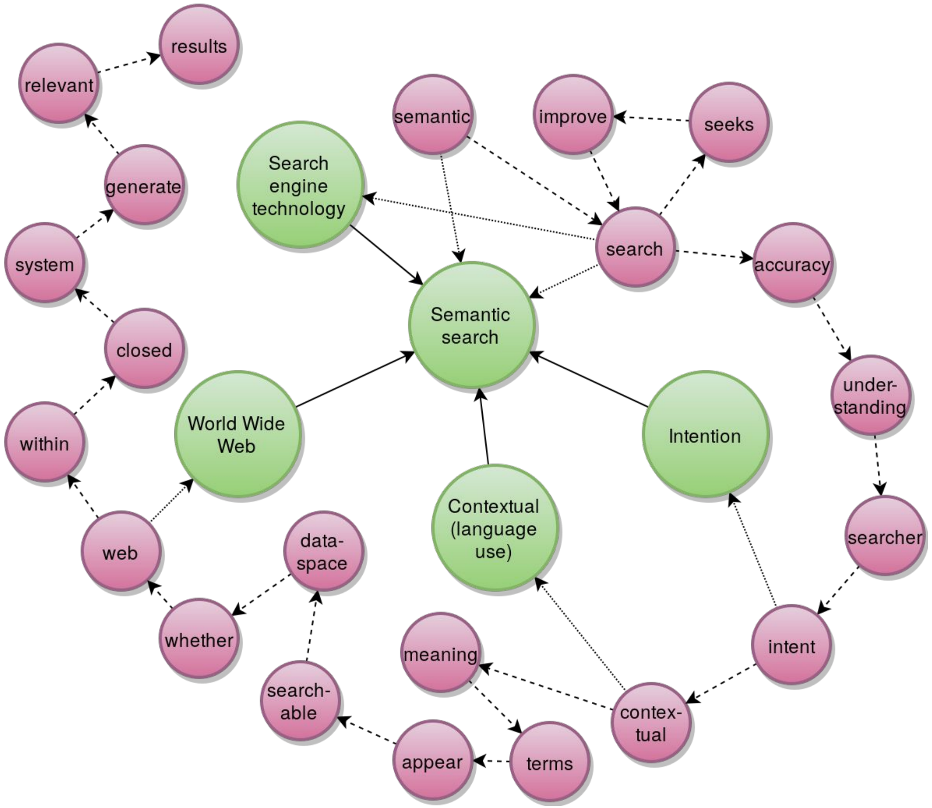
Graph-of-Entity

A baseline model for combined data.

Graph-of-Entity

Representation

- Collection-based graph.
- Each term links to a term that follows it.
- And to the entities that it might describe.
- Entities are linked according to the relations in the knowledge base.
- The graph is the index.



Graph-of-Entity

Ranking

- Ranking is computed based on the entity weight (EW):

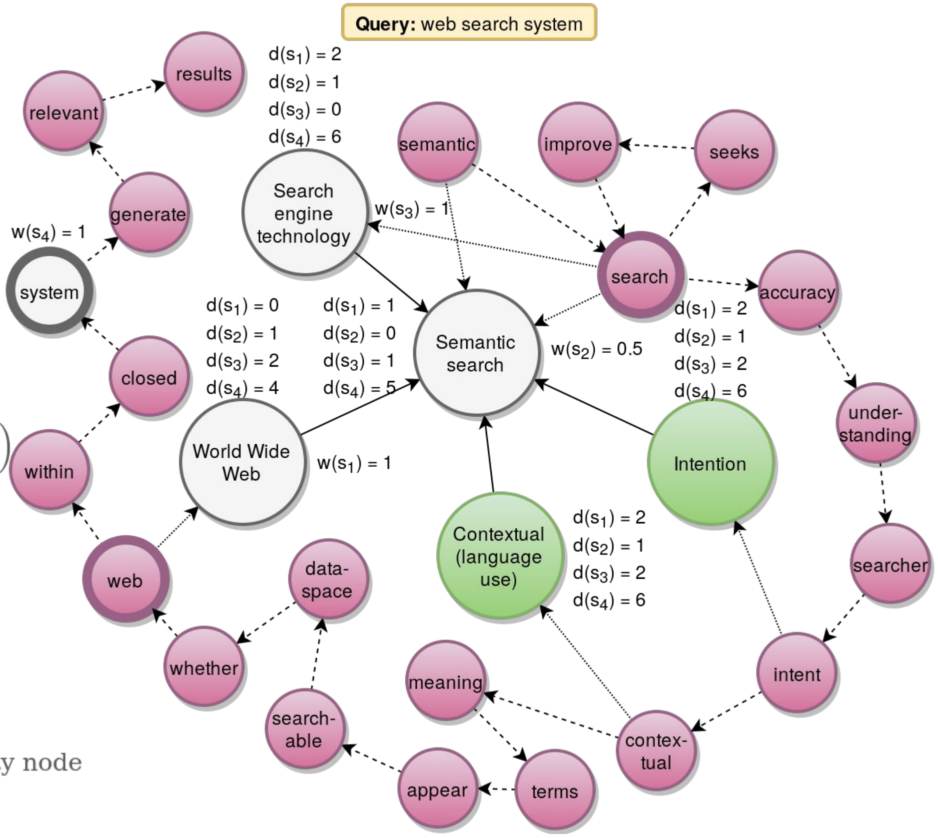
$$EW(e,q) = c(e,S_q) \times \frac{1}{|S_q|}\sum_{s\in S_q}\left(\frac{1}{|P_{es}|}\sum_{p_{es}\in P_{es}} w(s)\frac{1}{\epsilon(p_{es})}\right)$$

- Which considers coverage $c(e, S_q)$:

$$c(e,S_q) = \frac{|\{s\in S_q | \exists p_{es} \in P_{es}\}|}{|S_q|}$$

- And seed weight $w(s)$:

$$w(s) = \begin{cases} \frac{|\{e_{ts} \in E(G_e) | \forall t \exists q_n (t = q_n)\}|}{|\{e_{ts} \in E(G_e)\}|} & \text{if } s \text{ is an entity node} \\ 1 & \text{otherwise} \end{cases}$$



Graph-of-Entity

Ranking: coverage

- The coverage $c(e, S_q)$ measures the fraction of seeds that are connected to the entity to be scored:

$$c(e, S_q) = \frac{|\{s \in S_q \,|\, \exists p_{es} \in P_{es}\}|}{|S_q|}$$

Graph-of-Entity

Ranking: seed weight

- The seed weight $w(s)$ measures the "goodness" of a seed node in representing the query.
- It works like a degree of certainty analogous to a step in entity linking.
- An entity seed node is neighbor to one or more query term nodes. Its weight is the fraction of edges linking to query term nodes over the total number of edges (i.e., its degree).
- A term seed node is always a query term node and therefore has maximum weight.

$$w(s) = \begin{cases} \dfrac{|\{e_{ts} \in E(G_e) | \forall t \exists q_n (t = q_n)\}|}{|\{e_{ts} \in E(G_e)\}|} & \text{if } s \text{ is an entity node} \\ 1 & \text{otherwise} \end{cases}$$

Graph-of-Entity

Ranking: entity weight

- The entity weight $EW(e, q)$ scores an entity e, according to query q.
- It measures the proximity between the seed nodes s, representing the query, and an entity e.
- As an average of the weighted inverse length of the path, for all simple paths between e and s.
- This is then averaged over all seed nodes and boosted by coverage.

$$EW(e,q) = c(e, S_q) \times \frac{1}{|S_q|} \sum_{s \in S_q} \left(\frac{1}{|P_{es}|} \sum_{p_{es} \in P_{es}} w(s) \frac{1}{\epsilon(p_{es})} \right)$$

Evaluation

INEX 2009 Wikipedia Collection

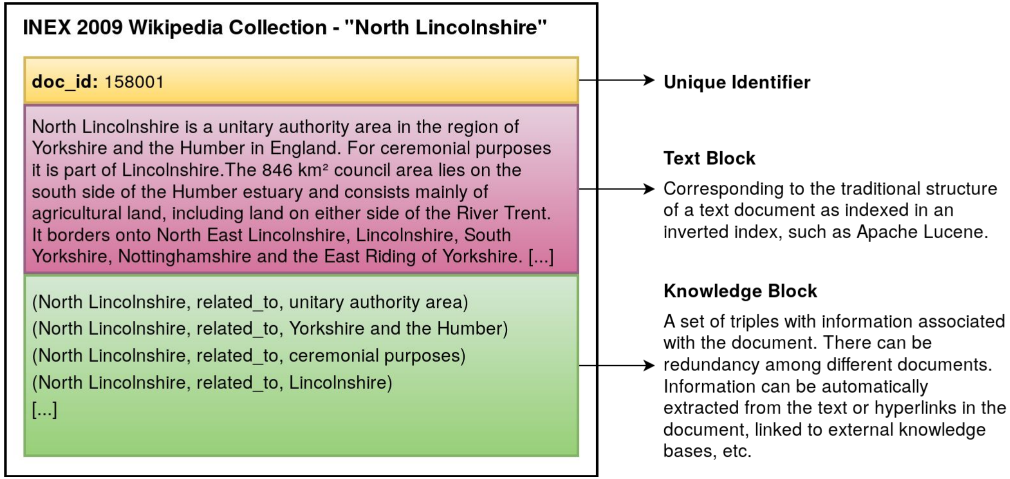
- Wikipedia XML corpus with 2.6 million articles.
- Semantically annotated based on 5,800 entity classes from the YAGO ontology.
- Snapshot from October 8, 2008.
- It's combined data!

- Task: ad hoc document retrieval, leveraging entities.
- Evaluation using topics and relevance judgments from INEX 2010 Ad Hoc track.
- Based on a sample of 10 topics, including all 7,487 documents mentioned in the relevance judgments.

INEX 2009 Wikipedia Collection

Extended document:

- Stripping text from XML provides a text block.
- Links between articles provide a knowledge block.



It's more than an entity-annotated document, since it might contain triples with external knowledge that extend the document beyond the immediate neighborhood of its entities.

Model	P@10	MAP	NDCG@10	Prec.	Recall
GoW	**0.3000**	**0.2333**	**0.3265**	0.1085	**0.9816**
GoE	0.1500	0.0399	0.1480	**0.1771**	0.2233

Topic ID	Topic Title (Query)	Average Precision	
		GoW	GoE
2010038	`[ dinosaur ]`	**0.6189**	0.0069
2010057	`[ Einstein Relativity theory ]`	0.2899	**0.1364**
2010003	`[ Monuments of India ]`	0.2888	**0.0000**
2010079	`[ famous chess endgames ]`	0.2541	0.0448
2010023	`[ retirement age ]`	0.2513	0.0027
2010040	`[ President of the United States ]`	0.2408	0.0051
2010096	`[ predictive analysis +logistic +regression model program application ]`	0.2185	0.0410
2010049	`[ European fruit trees ]`	0.0756	0.0119
2010014	`[ composer museum ]`	0.0624	0.1185
2010032	`[ japanese ballerina ]`	**0.0331**	0.0315
	MAP	0.2333	0.0399

- As it stands, the graph-of-entity (GoE) is, overall, less effective than the graph-of-word (GoW).
- GoE was only able to surpass GoW for topic 2010014: `[ composer museum ]`.

Conclusions

- We proposed a graph-based model for indexing and searching over combined data.
- We focused on a collection-based graph, as opposed to a document-based graph.
- The goal was to retain text-based properties, while integrating with a knowledge base.
- And using the graph as the index data structure.

Conclusions

- We expected that using a collection-based graph would result in improved retrieval effectiveness, as well as a way to naturally disambiguate entities.
- However, we obtained a significantly lower MAP for graph-of-entity, when compared to graph-of-word.

On a positive note...

- We were able to establish a graph-based strategy to jointly represent combined data, taking into account terms, entities and their relations in order to perform ranking.
- At the same time, we explored the consolidation of entity linking and entity ranking as a single ranking task over the graph-of-entity.
- While our proposed model was quite preliminary, it serves to illustrate the opportunity for research in unified frameworks that maximize information usage and exploit cross-referencing.

Future work

- Compare graph-of-word and graph-of-entity using sliding windows of equal size (i.e., consider more than just the following term in the graph-of-entity).
- Further explore available entity annotations in the INEX 2009 Wikipedia collection.
- Further improve the entity linking process and its integration in the ranking function.
- Tackle scalability issues, reducing the number of nodes and edges, or considering graph embedding approaches as an alternative.

Thank you!

You can experiment with the graph-of-entity and other retrieval models, like the hypergraph-of-entity (its successor), 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