Graph-Based Entity-Oriented Search

José Devezas <joseluisdevezas@gmail.com>
Thesis supervisor: Sérgio Nunes
INESC TEC and Faculty of Engineering, University of Porto

Thesis submitted to Faculty of Engineering of the University of Porto for the Doctor Degree in Computer Science within the Joint Doctoral Program in Computer Science of the Universities of Minho, Aveiro and Porto.

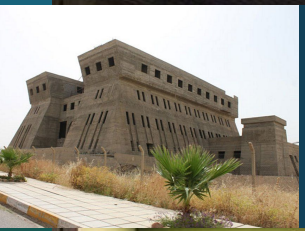
Porto, Portugal – January 26, 2021

REPÚBLICA PORTUGUESA | CIÊNCIA, TECNOLOGIA E ENSINO SUPERIOR
FCT Fundação para a Ciência e a Tecnologia
POCH
PORTUGAL 2020
UNIÃO EUROPEIA Fundo Social Europeu
INESCTEC
Universidade do Minho
universidade de aveiro
U.PORTO
MAPi DOCTORAL PROGRAMME IN COMPUTER SCIENCE

Introduction

For centuries, information has been organized, stored, and retrieved

- Clay tablets in Ashurbanipal
- Books in modern libraries
- Digitally encoded documents in computers
- Entities and their relations in knowledge bases



Library of Ashurbanipal (The British Museum)

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

– Krisztian Balog, 2018

- Entities and their relations
- Documents mentioning entities

- Knowledge bases
- Corpora

- Triplestores
- Inverted indexes

- Triplestores

Opportunity for a unified framework

- Inverted indexes

- Structured data and queries

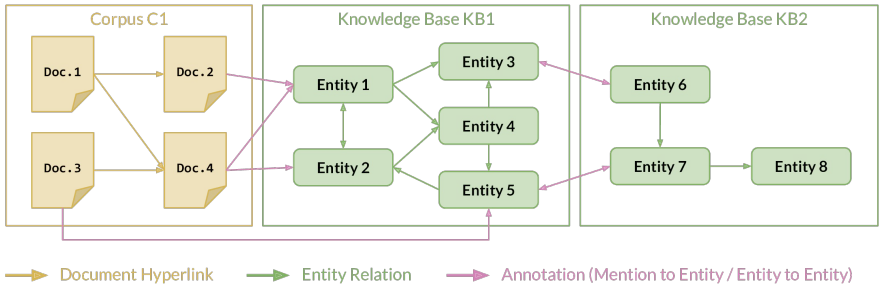
Opportunity for a unified framework

- Unstructured data and queries

A unified model for entity-oriented search

Combined data

- Text
- Entities
- Relations



A unified model for entity-oriented search

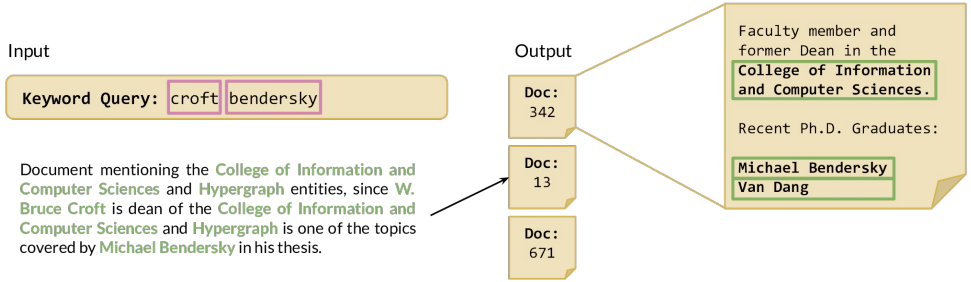
Retrieval tasks

- Ad hoc document retrieval
- Ad hoc entity retrieval
- Related entity finding
- Entity list completion

A unified model for entity-oriented search

Retrieval tasks

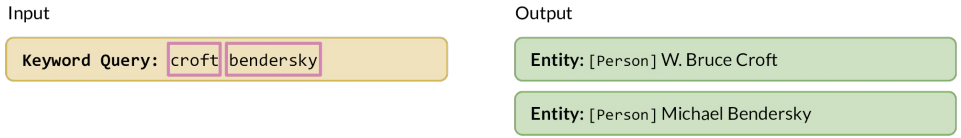
- Ad hoc document retrieval
- Ad hoc entity retrieval
- Related entity finding
- Entity list completion



A unified model for entity-oriented search

Retrieval tasks

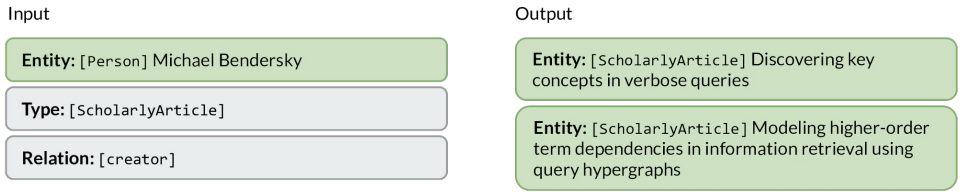
- Ad hoc document retrieval
- Ad hoc entity retrieval
- Related entity finding
- Entity list completion



A unified model for entity-oriented search

Retrieval tasks

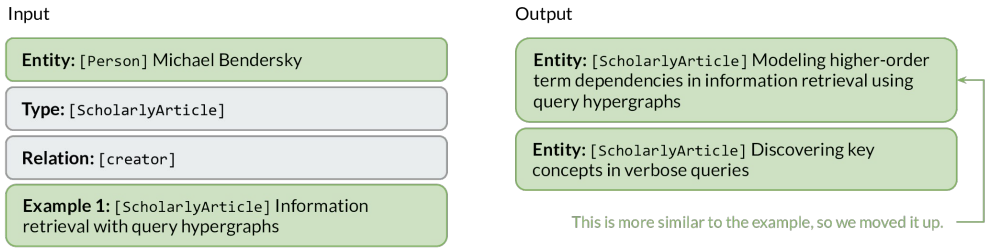
- Ad hoc document retrieval
- Ad hoc entity retrieval
- Related entity finding
- Entity list completion



A unified model for entity-oriented search

Retrieval tasks

- Ad hoc document retrieval
- Ad hoc entity retrieval
- Related entity finding
- Entity list completion



THESIS STATEMENT

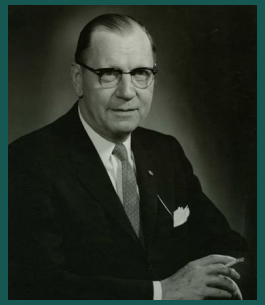
Graphs can be used to jointly index corpora and knowledge bases, supporting retrieval for multiple entity-oriented search tasks.

Main objectives

- Joint representation of terms, entities, and their relations
- Universal ranking function for multiple entity-oriented search tasks
- Improved retrieval effectiveness through the unification of information sources

State of the art

A breadth-first search for intersecting concepts in the worlds of graphs, entities, and documents.



Term Frequency

The weight of a term that occurs in a document is simply proportional to the term frequency.

– Hans Peter Luhn, 1957

Inverse Document Frequency

The specificity of a term can be quantified as an inverse function of the number of documents in which it occurs.

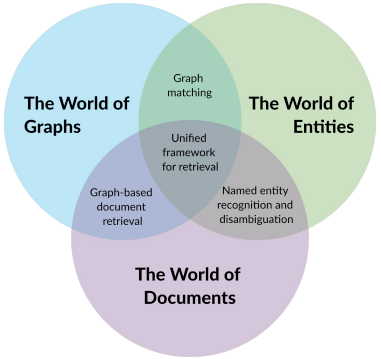
– Karen Spärck Jones, 1972

Inverted Files

The first important class of techniques for secondary key retrieval is based on the idea of an inverted file. This does not mean that the file is turned upside down; it means that the roles of records and attributes are reversed. Instead of listing the attributes of a given record, we list the records having a given attribute.

– Donald Ervin Knuth, 1973

Three axes covered



Classical models

- Virtual documents
- Triplestores
- Combined signals (single task or chained tasks)
- Joint indexing of text and triples (very few contributions)

Learning to rank

- Entity profiles represented as virtual documents
- Entity features
- Joint learning of word and entity representations

Graph-based models

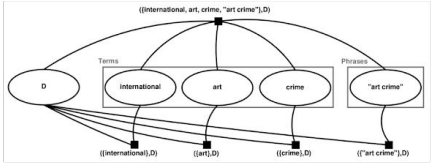
- Link analysis
- Text as a graph
- Knowledge graphs
- Entity graph from text
- Entity graph as a tensor
- Graph matching
- Hypergraph-based
- Random walk based

Anchor / core references

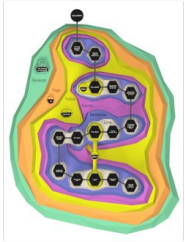
2007 2010 2012 2013 2017 2018

"73-87% of all queries contain entities"

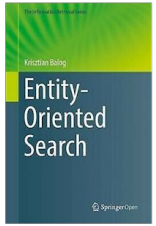
Concordance-Based Entity-Oriented Search
Bautin, M. and Skiena, S.



Modeling Higher-order Term Dependencies in Information Retrieval Using Query Hypergraphs
Bendersky, M. and Croft, W. B.

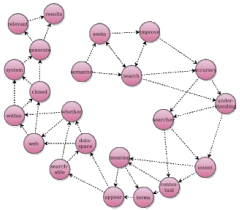


It's more than just overlap: Text As Graph
Dekker, R. H. and Birnbaum, D.



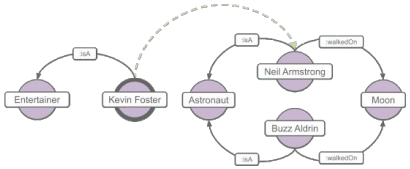
Entity-Oriented Search
Balog, K.

Ad-hoc object retrieval in the web of data
Pound, J., Mika, P. and Zaragoza, H.



Index for Efficient Semantic Full-text Search
Bast, H. and Buchhold, B.

Graph-of-word and TW-IDF: new approach to ad hoc IR
Rousseau, F. and Vazirgiannis, M.



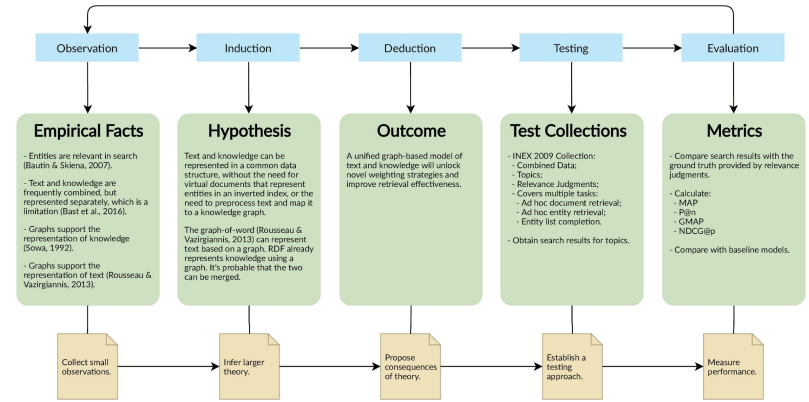
II

Materials and methods

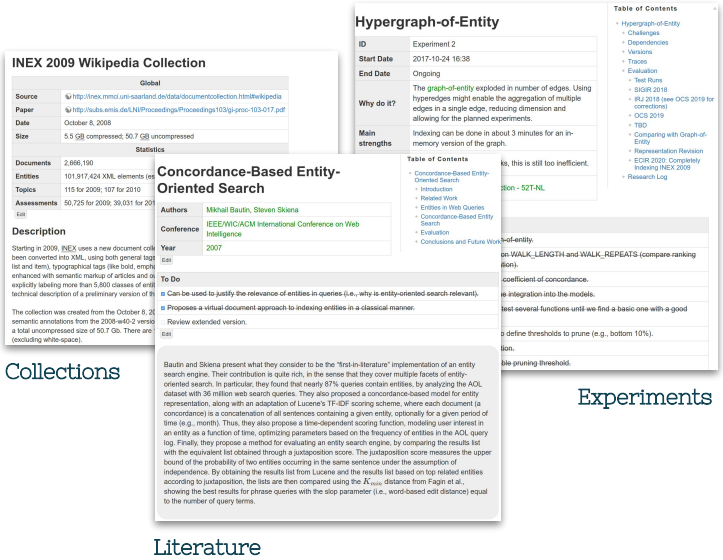
Empirical research supported on test collections and software

Research methodology

Empirical research



Systematic documentation



Test collections

INEX 2009 Wikipedia collection

- 2.6M XML documents
- Relevance judgments
  - 2010 Ad Hoc track
    - Document ranking
  - 2009 XER track
    - Entity ranking
    - List completion

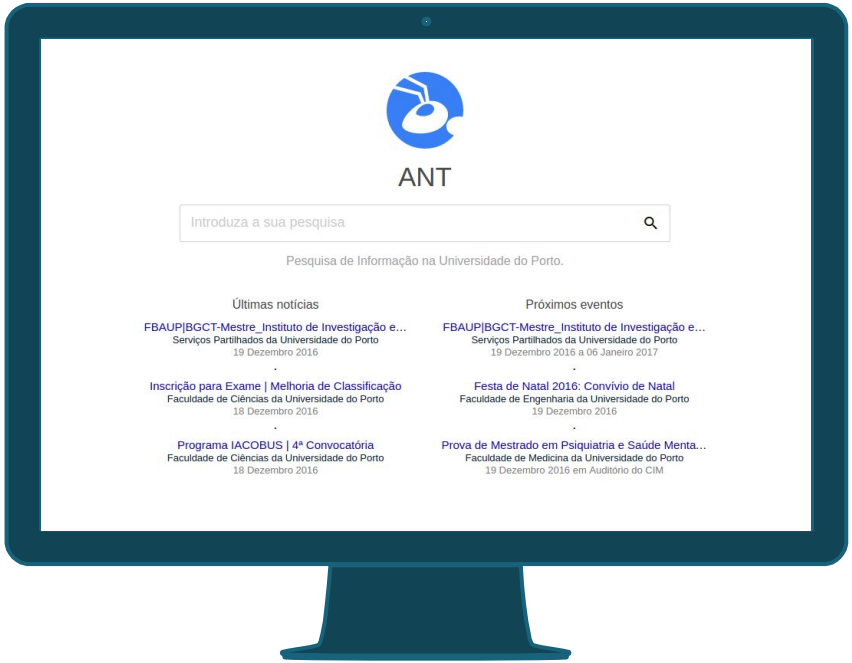
TREC Washington Post Corpus

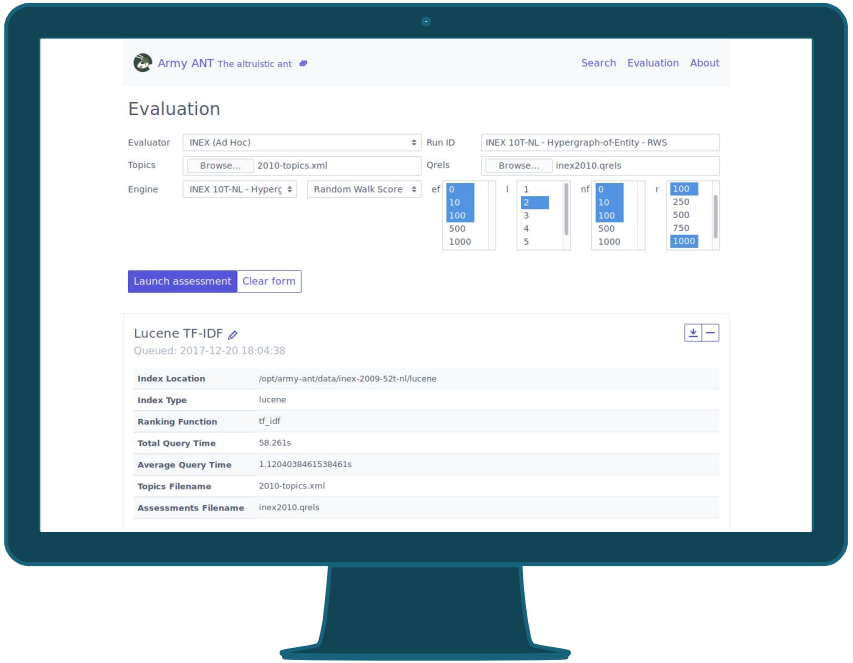
- 595K JSON documents
- Relevance judgements
  - 2018 Common Core track

TREC 2017 OpenSearch SSOAR

- 32K JSON documents
- Online evaluation
  - Team-draft interleaving
- Provided via Living Labs API

Software


ANT
Introduza a sua pesquisa
Pesquisa de Informação na Universidade do Porto.
Últimas notícias
FBAUP|BGCT-Mestre_Instituto de Investigação e...
Serviços Partilhados da Universidade do Porto
19 Dezembro 2016
Inscrição para Exame | Melhoria de Classificação
Faculdade de Ciências da Universidade do Porto
18 Dezembro 2016
Programa IACOBUS | 4ª Convocatória
Faculdade de Ciências da Universidade do Porto
18 Dezembro 2016
Próximos eventos
FBAUP|BGCT-Mestre_Instituto de Investigação e...
Serviços Partilhados da Universidade do Porto
19 Dezembro 2016 a 06 Janeiro 2017
Festa de Natal 2016: Convívio de Natal
Faculdade de Engenharia da Universidade do Porto
19 Dezembro 2016
Prova de Mestrado em Psiquiatria e Saúde Menta...
Faculdade de Medicina da Universidade do Porto
19 Dezembro 2016 em Auditório do CIM


Army ANT The altruistic ant
Search
Evaluation
About
Evaluation
Evaluator
INEX (Ad Hoc)
Run ID
INEX 10T-NL - Hypergraph-of-Entity - RWS
Topics
Browse...
2010-topics.xml
Qrels
Browse...
inex2010.qrels
Engine
INEX 10T-NL - Hyperç
Random Walk Score
ef
0
10
100
500
1000
l
1
2
3
4
5
nf
0
10
100
500
1000
r
100
250
500
750
1000
Launch assessment
Clear form
Lucene TF-IDF
Queued: 2017-12-20 18:04:38
Index Location
/opt/army-ant/data/inex-2009-52t-nl/lucene
Index Type
lucene
Ranking Function
tf_idf
Total Query Time
58.261s
Average Query Time
1.1204038461538461s
Topics Filename
2010-topics.xml
Assessments Filename
inex2010.qrels

III

Contributions

- Graph-of-entity
- Hypergraph-of-entity

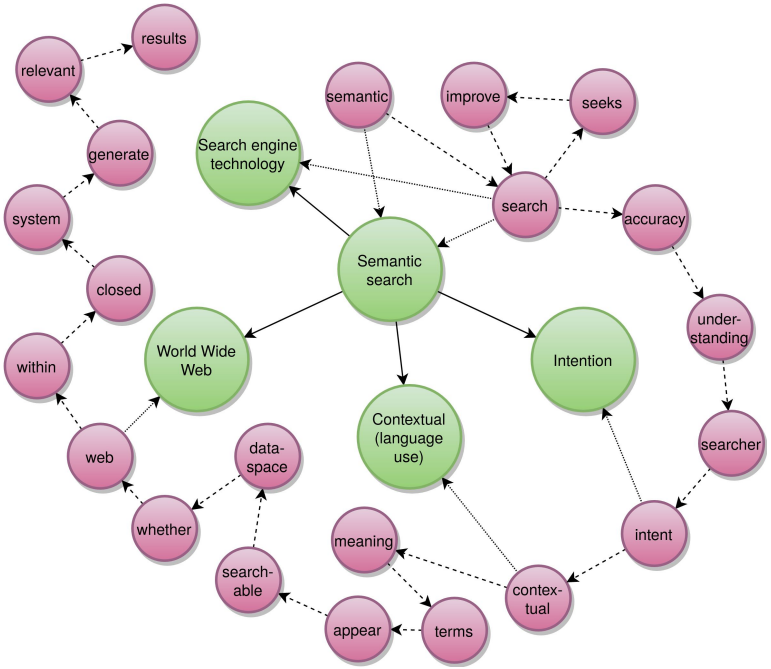
Example document

Semantic search seeks to improve search [Search Engine Technology] accuracy by understanding the searcher's intent [Intention] and the contextual [Contextual (language use)] meaning of terms as they appear in the searchable dataspace, whether on the Web [World Wide Web] or within a closed system, to generate more relevant results.

– 'Semantic search', Wikipedia, 9:10am, January 7, 2016

Graph-of-entity: Representation

- Nodes
  - ‘term’
  - ‘entity’
- Edges (directed and unweighted)
  - ‘before’
  - ‘related_to’
  - ‘contained_in’



Note: The direction for ‘related_to’ edges has been corrected.

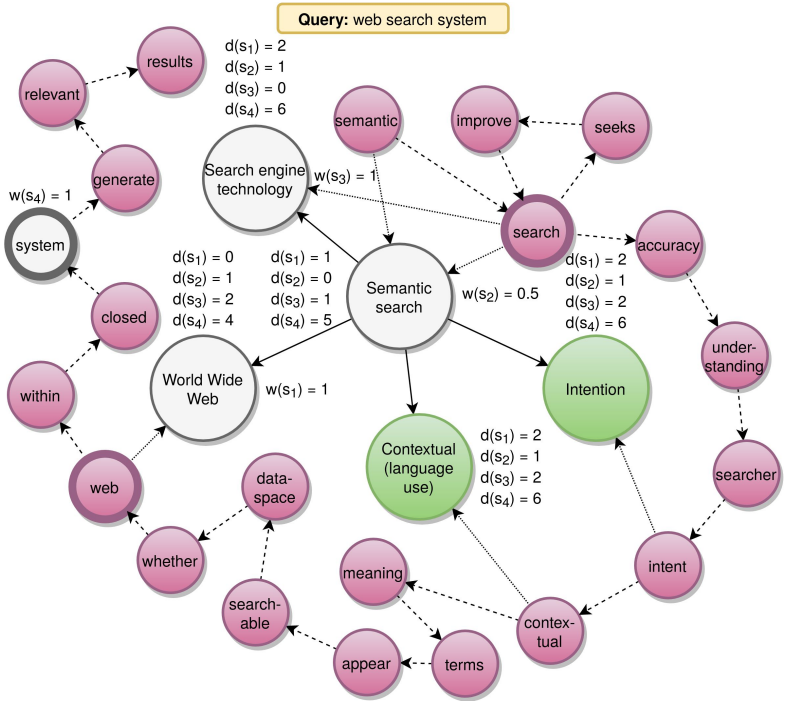
Seed nodes

- Map the query to the graph
- Can be expanded to adjacent entities
- Weighted according to their representability of the query

Graph-of-entity: Retrieval

Three components:

- Coverage
- Confidence weight
- Entity weight



Note: The direction for 'related_to' edges has been corrected.

Scaling issues

INEX 2009 Wikipedia subset

- 7,484 documents
- Graph-of-entity
  - 981,647 nodes
  - 9,942,647 edges

Scaling issues

INEX 2009 Wikipedia subset

- 7,484 documents
- Graph-of-entity
  - 981,647 nodes
  - 9,942,647 edges

Scaling issues

INEX 2009 Wikipedia subset

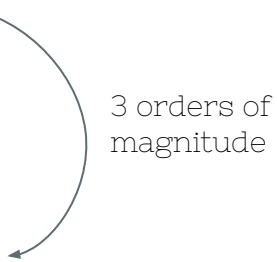
- 7,484 documents
- Graph-of-entity
  - 981,647 nodes
  - 9,942,647 edges

2 orders of magnitude

Scaling issues

INEX 2009 Wikipedia subset

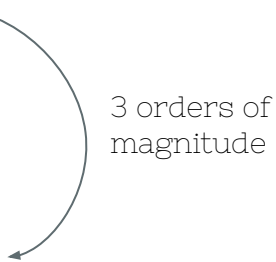
- 7,484 documents
- Graph-of-entity
  - 981,647 nodes
  - 9,942,647 edges



Scaling issues

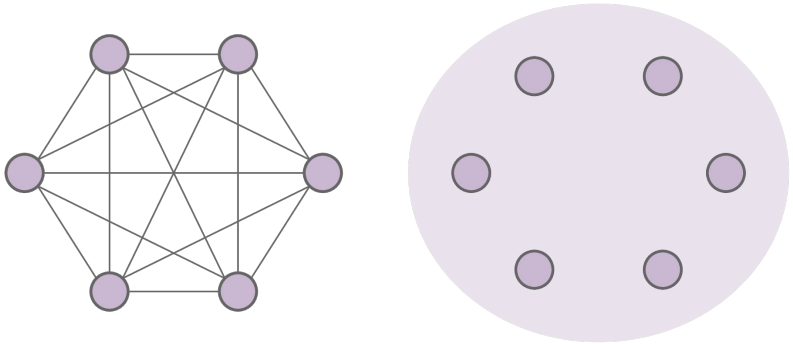
INEX 2009 Wikipedia subset

- 7,484 documents
- Graph-of-entity
  - 981,647 nodes
  - 9,942,647 edges

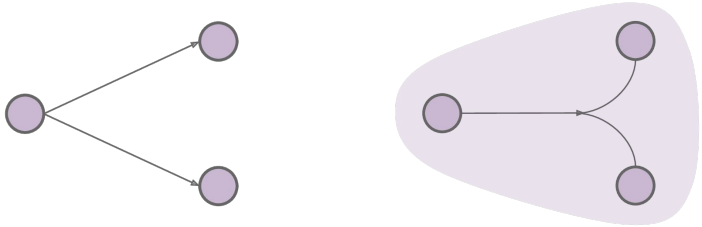


How could we reduce the number of edges per node?

From graphs to hypergraphs



Representing full connectivity
(e.g., synonyms)

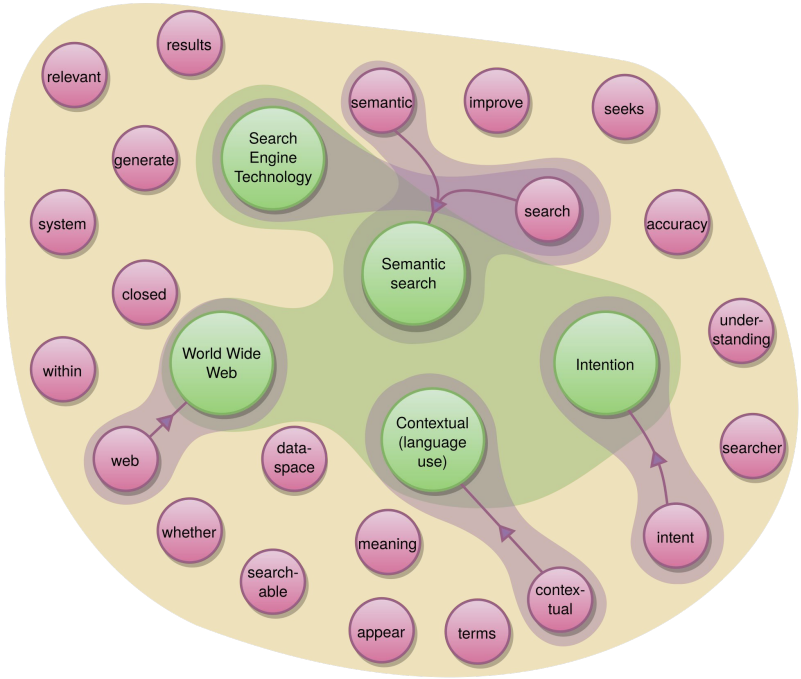


Representing directed n-ary connectivity
(e.g., e-mail message)

Hypergraph-of-entity: Joint representation model

Base model

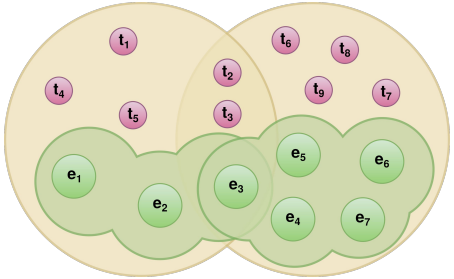
- Nodes
  - ‘term’
  - ‘entity’
- Hyperedges
  - ‘document’
  - ‘related_to’
  - ‘contained_in’



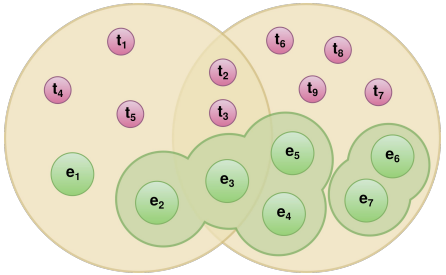
Hypergraph-of-entity: Joint representation model

Two types of ‘related_to’ hyperedges:

- Grouped by co-occurrence
  - Reinforces traversals across entities within the same document
- Grouped by subject
  - Reinforces traversals across entities within a context established by a subject entity



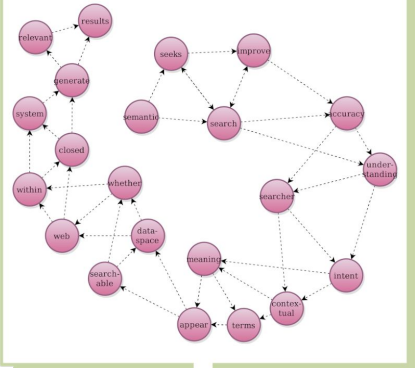
‘related_to’: Grouped by co-occurrence



‘related_to’: Grouped by subject

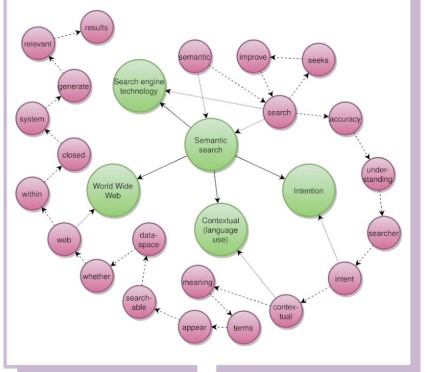
Scaling issues: Mitigated

Graph-of-word



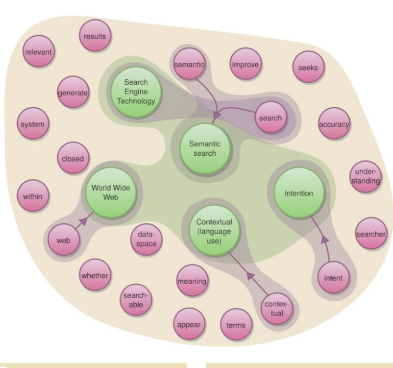
- 7,487 documents
- 492,185 vertices
- 22,906,803 edges
- |E| = 46.5 x |V|

Graph-of-entity



- 7,487 documents
- 981,647 vertices
- 9,942,647 edges
- |E| = 10.1 x |V|

Hypergraph-of-entity



- 7,487 documents
- 607,213 vertices
- 253,154 hyperedges
- |E| = 0.4 x |V|

Hypergraph-of-entity: Joint representation model

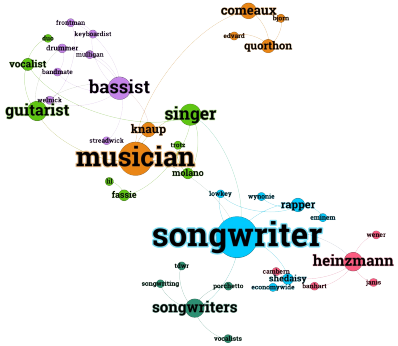
Extensions

Synonyms

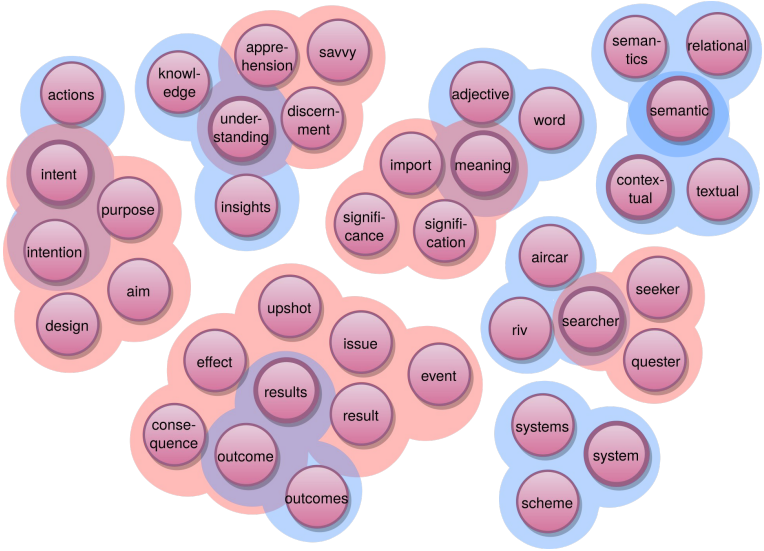
- WordNet 3.0 Nouns

Context

- Word2vec SimNet
  - size=100, window=5
  - 2-NN, cosine similarity > 0.5
  - Hyperedge per term neighborhood



Word2vec SimNet: Ego Network for 'musician', with depth 3.



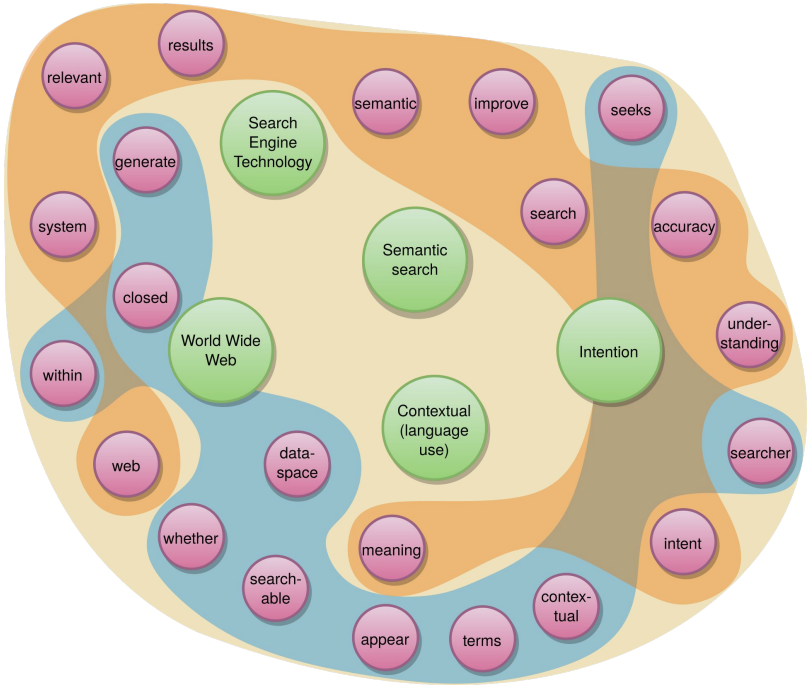
Extensions: Synonyms + Context

Hypergraph-of-entity: Joint representation model

Extensions

TF-bins (low TF / high TF):

- Discretization of term frequency
- Optionally weighted by percentile order, e.g.:
  - For P_4 = {25, 50, 75, 100}
  - w_{25}=1/4, w_{50}=2/4, w_{75}=3/4, w_{100}=1



Extensions: TF-bins (bin width = 2)

Hypergraph-of-entity: Joint representation model

Extensions

Weights

- Adds information that further constraints or guides the ranking function
- Introducing bias that affects both node and hyperedge sampling

Hypergraph-of-entity: Universal ranking function

Random walk score

- Biased random walks on a hypergraph
- Launched from each seed node
- Final score computed as:
  - ∑ weighted sum
    - Confidence weight
    - × visitation probability
  - × coverage

Hypergraph-of-entity: Universal ranking function

Table 7.4: Mapping entity-oriented search tasks to the hypergraph-of-entity.

	Query	Input	Results	Output
Ad hoc document retrieval	Keyword	Term nodes	Documents	Hyperedge ranking
Ad hoc entity retrieval	Keyword	Term nodes	Entities	Node ranking
Related entity finding	Entity	One entity node	Entities	Node ranking
Entity list completion	Entity	Multiple entity nodes	Entities	Node ranking

Hypergraph-of-entity: Universal ranking function

Table 7.5: Random walk score parameters and chosen configuration.

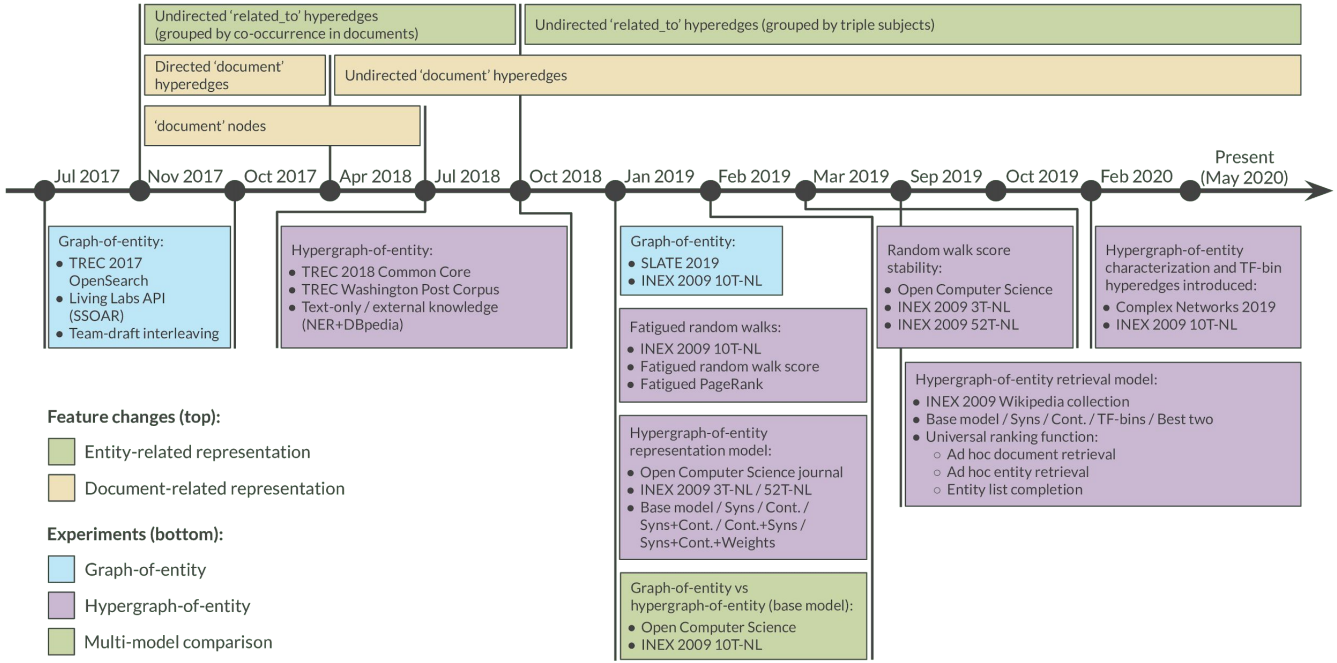
Parameter	Description	Configuration
ℓ	Length of the random walk.	2
r	Number of repeated random walks per seed node.	10,000
Δ_{nf}	Number of cycles of node fatigue (see Section B.2).	0
Δ_{ef}	Number of cycles of (hyper)edge fatigue (see Section B.2).	0
expansion	Whether to expand query to neighboring entities.	*false*
directed	Whether to consider or ignore direction.	*true*
weighted	Whether to consider node and hyperedge weights.	*false*

Hypergraph-of-entity: Characterization

- One of the few in-depth analyses of hypergraphs in applied network science
- Basic statistics over time (i.e., as index grows)
- Random walks and sampling used to reduce complexity
- Density based on a corresponding bipartite graph
  - Hyperedge-cardinality notation recognized as useful by the community

$$D = \frac{2\sum_k k|E_U^k| + \sum_{k_1,k_2}(k_1+k_2)|E_D^{k_1,k_2}|}{2(n+m)(n+m-1)}$$

Evaluation: Experimentation timeline



Evaluation: Main experiments

Retrieval performance over different representations:

- Text-only
- Base model
- Synonyms
- Context
- Syns+Cont.
- Cont.+Syns
- Syns+Cont.+Weights

Over different ranking function parameter configurations:

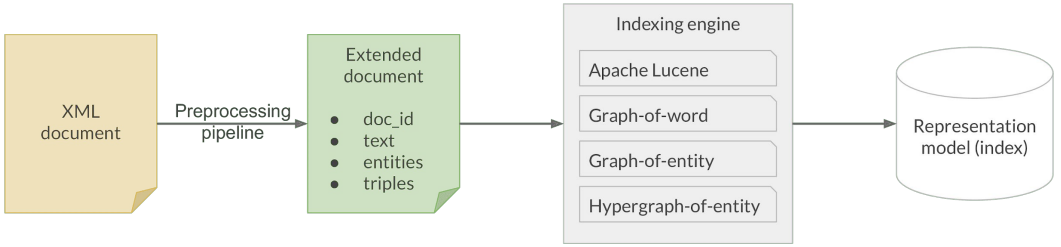
- Best results for:
  - Low ℓ
  - High r
  - No fatigue
- Variable results for:
  - Expansion
  - Weights

And for multiple tasks over the same index.

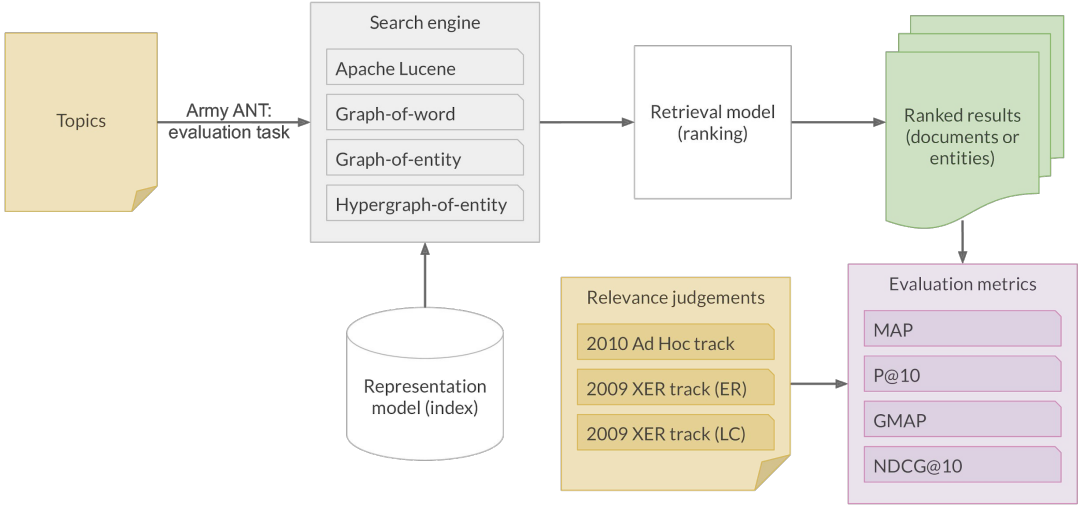
Evaluation: INEX 2009 Wikipedia collection

Indexing, retrieval and ranking, and evaluation

Indexing



Ranked retrieval & evaluation



Evaluation: INEX 2009 Wikipedia collection

Graph-of-entity

- Evaluation based on the 2010 Ad Hoc track qrels
- Ranking of entity nodes with an associated document

Hypergraph-of-entity (subsets-only):

- Evaluation based on the 2010 Ad Hoc track qrels
- Ranking of document hyperedges

Evaluation: INEX 2009 10T-NL Wikipedia subset

Ad hoc document retrieval

Graph-of-entity vs Hypergraph-of-entity

Table 9.5: Graph-of-entity (GoE) vs hypergraph-of-entity (HGoE) with $\ell = 2$.

(a) Effectiveness (highest values for Lucene and graph-based models in bold).

Index	Ranking	GMAP	MAP	Precision	Recall	NDCG@10	P@10
Lucene	TF-IDF	0.1540	0.1710	0.1389	0.8007	0.2671	0.2800
	BM25	**0.2802**	**0.2963**	**0.1396**	**0.8241**	**0.5549**	**0.5000**
GoE	EW	0.0003	0.0399	**0.1771**	0.2233	0.1480	0.1500
HGoE	RWS($r = 10^1$)	0.0000	0.0485	0.0734	0.3085	0.1229	0.1200
	RWS($r = 10^2$)	0.0546	0.1118	0.0342	0.7554	0.1474	0.1500
	RWS($r = 10^3$)	0.1017	0.1492	0.0199	0.9122	**0.2074**	**0.2200**
	RWS($r = 10^4$)	**0.1224**	**0.1689**	0.0167	**0.9922**	0.1699	0.1700

(b) Efficiency (lowest times for Lucene and graph-based models in bold).

Index	Ranking	Indexing Time (Total)	Search Time (Avg./Query)	Nodes	Edges
Lucene	TF-IDF	**27s 769ms**	**209ms**	N/A	N/A
	BM25		316ms		
GoE	EW	1h 38m	21s 557ms	981,647	9,942,647
HGoE	RWS($r = 10^1$)	**53s 922ms**	**943ms**	607,213	253,154
	RWS($r = 10^2$)		11s 134ms		
	RWS($r = 10^3$)		1m 17s 540ms		
	RWS($r = 10^4$)		13m 04s 057ms		

Evaluation: INEX 2009 10T-NL Wikipedia subset

Ad hoc document retrieval

Graph-of-entity vs Hypergraph-of-entity

Table 9.5: Graph-of-entity (GoE) vs hypergraph-of-entity (HGoE) with $\ell = 2$.

(a) Effectiveness (highest values for Lucene and graph-based models in bold).

Index	Ranking	GMAP	MAP	Precision	Recall	NDCG@10	P@10
Lucene	TF-IDF	0.1540	**0.1710**	0.1389	0.8007	0.2671	0.2800
	BM25	**0.2802**	**0.2963**	**0.1396**	**0.8241**	**0.5549**	**0.5000**
GoE	EW	0.0003	0.0399	**0.1771**	0.2233	0.1480	0.1500
HGoE	RWS($r = 10^1$)	0.0000	0.0485	0.0734	0.3085	0.1229	0.1200
	RWS($r = 10^2$)	0.0546	0.1118	0.0342	0.7554	0.1474	0.1500
	RWS($r = 10^3$)	0.1017	0.1492	0.0199	0.9122	**0.2074**	**0.2200**
	RWS($r = 10^4$)	**0.1224**	**0.1689**	0.0167	**0.9922**	0.1699	0.1700

(b) Efficiency (lowest times for Lucene and graph-based models in bold).

Index	Ranking	Indexing Time (Total)	Search Time (Avg./Query)	Nodes	Edges
Lucene	TF-IDF	**27s 769ms**	**209ms**	N/A	N/A
	BM25		316ms		
GoE	EW	1h 38m	21s 557ms	981,647	9,942,647
HGoE	RWS($r = 10^1$)	**53s 922ms**	**943ms**	607,213	253,154
	RWS($r = 10^2$)		11s 134ms		
	RWS($r = 10^3$)		1m 17s 540ms		
	RWS($r = 10^4$)		13m 04s 057ms		

Evaluation: INEX 2009 10T-NL Wikipedia subset

Ad hoc document retrieval

Graph-of-entity vs Hypergraph-of-entity

Table 9.5: Graph-of-entity (GoE) vs hypergraph-of-entity (HGoE) with $\ell = 2$.

(a) Effectiveness (highest values for Lucene and graph-based models in bold).

Index	Ranking	GMAP	MAP	Precision	Recall	NDCG@10	P@10
Lucene	TF-IDF	0.1540	0.1710	0.1389	0.8007	0.2671	0.2800
	BM25	**0.2802**	**0.2963**	**0.1396**	**0.8241**	**0.5549**	**0.5000**
GoE	EW	0.0003	0.0399	**0.1771**	0.2233	0.1480	0.1500
HGoE	RWS($r = 10^1$)	0.0000	0.0485	0.0734	0.3085	0.1229	0.1200
	RWS($r = 10^2$)	0.0546	0.1118	0.0342	0.7554	0.1474	0.1500
	RWS($r = 10^3$)	0.1017	0.1492	0.0199	0.9122	**0.2074**	**0.2200**
	RWS($r = 10^4$)	**0.1224**	**0.1689**	0.0167	**0.9922**	0.1699	0.1700

(b) Efficiency (lowest times for Lucene and graph-based models in bold).

Index	Ranking	Indexing Time (Total)	Search Time (Avg./Query)	Nodes	Edges
Lucene	TF-IDF	**27s 769ms**	**209ms**	N/A	N/A
	BM25		316ms		
GoE	EW	1h 38m	21s 557ms	981,647	9,942,647
HGoE	RWS($r = 10^1$)	**53s 922ms**	**943ms**	607,213	253,154
	RWS($r = 10^2$)		11s 134ms		
	RWS($r = 10^3$)		1m 17s 540ms		
	RWS($r = 10^4$)		13m 04s 057ms		

Evaluation: INEX 2009 10T-NL Wikipedia subset

Ad hoc document retrieval

Graph-of-entity vs Hypergraph-of-entity

Table 9.5: Graph-of-entity (GoE) vs hypergraph-of-entity (HGoE) with $\ell = 2$.

(a) Effectiveness (highest values for Lucene and graph-based models in bold).

Index	Ranking	GMAP	MAP	Precision	Recall	NDCG@10	P@10
Lucene	TF-IDF	0.1540	0.1710	0.1389	0.8007	0.2671	0.2800
	BM25	**0.2802**	**0.2963**	**0.1396**	**0.8241**	**0.5549**	**0.5000**
GoE	EW	0.0003	0.0399	**0.1771**	0.2233	0.1480	0.1500
HGoE	RWS($r = 10^1$)	0.0000	0.0485	0.0734	0.3085	0.1229	0.1200
	RWS($r = 10^2$)	0.0546	0.1118	0.0342	0.7554	0.1474	0.1500
	RWS($r = 10^3$)	0.1017	0.1492	0.0199	0.9122	**0.2074**	**0.2200**
	RWS($r = 10^4$)	**0.1224**	**0.1689**	0.0167	**0.9922**	0.1699	0.1700

(b) Efficiency (lowest times for Lucene and graph-based models in bold).

Index	Ranking	Indexing Time (Total)	Search Time (Avg./Query)	Nodes	Edges
Lucene	TF-IDF	27s 769ms	**209ms**	N/A	N/A
	BM25		316ms		
GoE	EW	1h 38m	21s 557ms	981,647	9,942,647
HGoE	RWS($r = 10^1$)	53s 922ms	**943ms**	607,213	253,154
	RWS($r = 10^2$)		11s 134ms		
	RWS($r = 10^3$)		1m 17s 540ms		
	RWS($r = 10^4$)		13m 04s 057ms		

Evaluation: INEX 2009 52T-NL Wikipedia subset

Ad hoc document retrieval

Comparing representation models

Table 9.4: Best overall parameter configuration according to the mean average precision.

(a) Effectiveness (highest values for Lucene and hypergraph-of-entity in bold; differences in MAP are not statistically significant, except between the Lucene baselines and the hypergraph-of-entity indexes).

Index	Ranking	GMAP	MAP	Precision	Recall	NDCG@10	P@10
Lucene	TF-IDF	0.1345	0.1689	**0.0650**	0.8476	0.2291	0.2346
	BM25	**0.2740**	**0.3269**	0.0647	**0.8598**	**0.5607**	**0.5250**
Hypergraph-of-Entity: Random Walk Score ($\ell = 2$, $r = 10^3$)							
Base Model	RWS	0.0285	0.0864	0.0219	0.8003	**0.1413**	**0.1269**
Syns	RWS	0.0281	0.0840	0.0225	0.8099	0.1301	0.1231
Context	RWS	0.0134	0.0811	0.0220	0.8027	0.1218	0.1192
Syns+Context	RWS	0.0299	0.0837	0.0236	0.8069	0.1310	0.1231
Context+Syns	RWS	0.0296	0.0814	0.0242	**0.8148**	0.1256	0.1250
Syns+Cont.+Weights	RWS	**0.0313**	**0.0884**	**0.0274**	0.8059	0.1256	0.1154

(b) Efficiency (lowest times for Lucene and hypergraph-of-entity in bold).

Index	Ranking	Indexing Time		Search Time	
		Avg./Doc	Total	Avg./Query	Total
Lucene	TF-IDF	2.16ms	1m 21s 382ms	**1s 148ms**	**59s 698ms**
	BM25			1s 220ms	1m 03s 461ms
Hypergraph-of-Entity: Random Walk Score ($\ell = 2$, $r = 10^3$)					
Base Model	RWS	6.52ms	4m 05s 612ms	**3m 22s 826ms**	**2h 55m 47s**
Syns	RWS	**6.22ms**	**3m 54s 587ms**	3m 31s 038ms	3h 02m 54s
Context	RWS	6.35ms	3m 59s 446ms	3m 35s 623ms	3h 06m 52s
Syns+Context	RWS	6.29ms	3m 57s 264ms	3m 33s 000ms	3h 04m 36s
Context+Syns	RWS	6.33ms	3m 58s 659ms	3m 36s 487ms	3h 07m 37s
Syns+Cont.+Weights	RWS	6.52ms	4m 05s 984ms	10m 55s 590ms	9h 28m 11s

Evaluation: INEX 2009 52T-NL Wikipedia subset

Ad hoc document retrieval

Comparing representation models

Table 9.4: Best overall parameter configuration according to the mean average precision.

(a) Effectiveness (highest values for Lucene and hypergraph-of-entity in bold; differences in MAP are not statistically significant, except between the Lucene baselines and the hypergraph-of-entity indexes).

Index	Ranking	GMAP	MAP	Precision	Recall	NDCG@10	P@10
Lucene	TF-IDF	0.1345	0.1689	**0.0650**	0.8476	0.2291	0.2346
	BM25	**0.2740**	**0.3269**	0.0647	**0.8598**	**0.5607**	**0.5250**
Hypergraph-of-Entity: Random Walk Score ($\ell = 2$, $r = 10^3$)							
Base Model	RWS	0.0285	0.0864	0.0219	0.8003	**0.1413**	**0.1269**
Syns	RWS	0.0281	0.0840	0.0225	0.8099	0.1301	0.1231
Context	RWS	0.0134	0.0811	0.0220	0.8027	0.1218	0.1192
Syns+Context	RWS	0.0299	0.0837	0.0236	0.8069	0.1310	0.1231
Context+Syns	RWS	0.0296	0.0814	0.0242	**0.8148**	0.1256	0.1250
Syns+Cont.+Weights	RWS	**0.0313**	**0.0884**	**0.0274**	0.8059	0.1256	0.1154

(b) Efficiency (lowest times for Lucene and hypergraph-of-entity in bold).

Index	Ranking	Indexing Time		Search Time	
		Avg./Doc	Total	Avg./Query	Total
Lucene	TF-IDF	2.16ms	1m 21s 382ms	**1s 148ms**	**59s 698ms**
	BM25			1s 220ms	1m 03s 461ms
Hypergraph-of-Entity: Random Walk Score ($\ell = 2$, $r = 10^3$)					
Base Model	RWS	6.52ms	4m 05s 612ms	**3m 22s 826ms**	**2h 55m 47s**
Syns	RWS	**6.22ms**	**3m 54s 587ms**	3m 31s 038ms	3h 02m 54s
Context	RWS	6.35ms	3m 59s 446ms	3m 35s 623ms	3h 06m 52s
Syns+Context	RWS	6.29ms	3m 57s 264ms	3m 33s 000ms	3h 04m 36s
Context+Syns	RWS	6.33ms	3m 58s 659ms	3m 36s 487ms	3h 07m 37s
Syns+Cont.+Weights	RWS	6.52ms	4m 05s 984ms	10m 55s 590ms	9h 28m 11s

Evaluation: INEX 2009 52T-NL Wikipedia subset

Ad hoc document retrieval

Comparing representation models

Table 9.4: Best overall parameter configuration according to the mean average precision.

(a) Effectiveness (highest values for Lucene and hypergraph-of-entity in bold; differences in MAP are not statistically significant, except between the Lucene baselines and the hypergraph-of-entity indexes).

Index	Ranking	GMAP	MAP	Precision	Recall	NDCG@10	P@10
Lucene	TF-IDF	0.1345	0.1689	**0.0650**	0.8476	0.2291	0.2346
	BM25	**0.2740**	**0.3269**	0.0647	**0.8598**	**0.5607**	**0.5250**
Hypergraph-of-Entity: Random Walk Score ($\ell = 2$, $r = 10^3$)							
Base Model	RWS	0.0285	0.0864	0.0219	0.8003	**0.1413**	**0.1269**
Syns	RWS	0.0281	0.0840	0.0225	0.8099	0.1301	0.1231
Context	RWS	0.0134	0.0811	0.0220	0.8027	0.1218	0.1192
Syns+Context	RWS	0.0299	0.0837	0.0236	0.8069	0.1310	0.1231
Context+Syns	RWS	0.0296	0.0814	0.0242	**0.8148**	0.1256	0.1250
Syns+Cont.+Weights	RWS	**0.0313**	**0.0884**	**0.0274**	0.8059	0.1256	0.1154

(b) Efficiency (lowest times for Lucene and hypergraph-of-entity in bold).

Index	Ranking	Indexing Time		Search Time	
		Avg./Doc	Total	Avg./Query	Total
Lucene	TF-IDF	2.16ms	1m 21s 382ms	**1s 148ms**	**59s 698ms**
	BM25			1s 220ms	1m 03s 461ms
Hypergraph-of-Entity: Random Walk Score ($\ell = 2$, $r = 10^3$)					
Base Model	RWS	6.52ms	4m 05s 612ms	**3m 22s 826ms**	**2h 55m 47s**
Syns	RWS	**6.22ms**	**3m 54s 587ms**	3m 31s 038ms	3h 02m 54s
Context	RWS	6.35ms	3m 59s 446ms	3m 35s 623ms	3h 06m 52s
Syns+Context	RWS	6.29ms	3m 57s 264ms	3m 33s 000ms	3h 04m 36s
Context+Syns	RWS	6.33ms	3m 58s 659ms	3m 36s 487ms	3h 07m 37s
Syns+Cont.+Weights	RWS	6.52ms	4m 05s 984ms	10m 55s 590ms	9h 28m 11s

Evaluation: INEX 2009 52T-NL Wikipedia subset

Ad hoc document retrieval

Comparing representation models

Table 9.4: Best overall parameter configuration according to the mean average precision.

(a) Effectiveness (highest values for Lucene and hypergraph-of-entity in bold; differences in MAP are not statistically significant, except between the Lucene baselines and the hypergraph-of-entity indexes).

Index	Ranking	GMAP	MAP	Precision	Recall	NDCG@10	P@10
Lucene	TF-IDF	0.1345	0.1689	**0.0650**	0.8476	0.2291	0.2346
	BM25	**0.2740**	**0.3269**	0.0647	**0.8598**	**0.5607**	**0.5250**
Hypergraph-of-Entity: Random Walk Score ($\ell = 2$, $r = 10^3$)							
Base Model	RWS	0.0285	0.0864	0.0219	0.8003	**0.1413**	**0.1269**
Syns	RWS	0.0281	0.0840	0.0225	0.8099	0.1301	0.1231
Context	RWS	0.0134	0.0811	0.0220	0.8027	0.1218	0.1192
Syns+Context	RWS	0.0299	0.0837	0.0236	0.8069	0.1310	0.1231
Context+Syns	RWS	0.0296	0.0814	0.0242	**0.8148**	0.1256	0.1250
Syns+Cont.+Weights	RWS	**0.0313**	**0.0884**	**0.0274**	0.8059	0.1256	0.1154

(b) Efficiency (lowest times for Lucene and hypergraph-of-entity in bold).

Index	Ranking	Indexing Time		Search Time	
		Avg./Doc	Total	Avg./Query	Total
Lucene	TF-IDF	2.16ms	1m 21s 382ms	**1s 148ms**	**59s 698ms**
	BM25			1s 220ms	1m 03s 461ms
Hypergraph-of-Entity: Random Walk Score ($\ell = 2$, $r = 10^3$)					
Base Model	RWS	6.52ms	4m 05s 612ms	**3m 22s 826ms**	**2h 55m 47s**
Syns	RWS	**6.22ms**	**3m 54s 587ms**	3m 31s 038ms	3h 02m 54s
Context	RWS	6.35ms	3m 59s 446ms	3m 35s 623ms	3h 06m 52s
Syns+Context	RWS	6.29ms	3m 57s 264ms	3m 33s 000ms	3h 04m 36s
Context+Syns	RWS	6.33ms	3m 58s 659ms	3m 36s 487ms	3h 07m 37s
Syns+Cont.+Weights	RWS	6.52ms	4m 05s 984ms	10m 55s 590ms	9h 28m 11s

Evaluation: INEX 2009 Wikipedia collection

Hypergraph-of-entity (full collection):

- 2010 Ad Hoc track qrels
  - Document ranking
- 2009 XER track qrels for:
  - Entity ranking
  - List completion
- Using keyword-based document profiles

Evaluation: INEX 2009 Wikipedia full collection

Multiple tasks

Keyword-based document profiles

Note: The entity index was created from sentence-based entity profiles

Table 10.6: Overall comparison of retrieval performance, for multiple entity-oriented search tasks, based on the complete INEX 2009 Wikipedia collection.

Index	Ranking	MAP	P@10	Index Time	Avg./Query	Nodes	Edges
AD HOC DOCUMENT RETRIEVAL							
Lucene	TF-IDF	0.0228	0.0692	15h 06m	460ms	–	–
	BM25	0.0324	0.1173		370ms	–	–
Fixed parameters over HGoE variations: RWS($\ell = 2, r = 10^4, \Delta_{nf} = 0, \Delta_{ef} = 0, exp = F, wei = F$)							
Base Model	RWS	0.0863	0.2462	33h 53m	23s 405ms	3,506,823	7,721,743
Syns	RWS	0.0937	0.2615	33h 05m	55s 555ms	3,510,462	7,734,931
Cont.	RWS	0.0869	0.2654	34h 37m	24s 348ms	3,604,185	7,929,841
TF-Bins$_2$	RWS	0.0172	0.0500	35h 26m	2m 58s	3,506,823	10,338,867
Syns+Cont.	RWS	0.0882	0.2692	37h 16m	23s 265ms	3,606,693	7,945,083
AD HOC ENTITY RETRIEVAL							
Lucene	TF-IDF	0.0373	0.0636	59h 17m	1s 370ms	–	–
	BM25	0.0668	0.1182		798ms	–	–
Fixed parameters over HGoE variations: RWS($\ell = 2, r = 10^4, \Delta_{nf} = 0, \Delta_{ef} = 0, exp = F, wei = F$)							
Base Model	RWS	0.1390	0.2455	33h 53m	26s 330ms	3,506,823	7,721,743
Syns	RWS	0.1337	0.2473	33h 05m	30s 232ms	3,510,462	7,734,931
Cont.	RWS	0.1304	0.2364	34h 37m	27s 620ms	3,604,185	7,929,841
TF-Bins$_2$	RWS	0.0300	0.1145	35h 26m	4m 41s	3,506,823	10,338,867
Syns+Cont.	RWS	0.1313	0.2509	37h 16m	26s 877ms	3,606,693	7,945,083
ENTITY LIST COMPLETION							
Lucene	TF-IDF	0.0558	0.1000	59h 17m	1s 230ms	–	–
	BM25	0.0666	0.1250		1s 221ms	–	–
Fixed parameters over HGoE variations: RWS($\ell = 2, r = 10^4, \Delta_{nf} = 0, \Delta_{ef} = 0, exp = F, wei = F$)							
Base Model	RWS	0.0879	0.0769	33h 53m	19s 162ms	3,506,823	7,721,743
Syns	RWS	0.0857	0.0635	33h 05m	19s 875ms	3,510,462	7,734,931
Cont.	RWS	0.0875	0.0692	34h 37m	19s 422ms	3,604,185	7,929,841
TF-Bins$_2$	RWS	0.0006	0.0058	35h 26m	1m 08s	3,506,823	10,338,867
Syns+Cont.	RWS	0.0884	0.0788	37h 16m	19s 824ms	3,606,693	7,945,083

Evaluation: INEX 2009 Wikipedia full collection

Multiple tasks

Keyword-based document profiles

Note: The entity index was created from sentence-based entity profiles

Table 10.6: Overall comparison of retrieval performance, for multiple entity-oriented search tasks, based on the complete INEX 2009 Wikipedia collection.

Index	Ranking	MAP	P@10	Index Time	Avg./Query	Nodes	Edges
Ad hoc document retrieval							
Lucene	TF-IDF	0.0228	0.0692	15h 06m	460ms	–	–
	BM25	0.0324	0.1173		370ms	–	–
Fixed parameters over HGoE variations: $RWS(\ell = 2, r = 10^4, \Delta_{nf} = 0, \Delta_{ef} = 0, exp = F, wei = F)$							
Base Model	RWS	0.0863	0.2462	33h 53m	23s 405ms	3,506,823	7,721,743
Syns	RWS	0.0937	0.2615	33h 05m	55s 555ms	3,510,462	7,734,931
Cont.	RWS	0.0869	0.2654	34h 37m	24s 348ms	3,604,185	7,929,841
TF-Bins$_2$	RWS	0.0172	0.0500	35h 26m	2m 58s	3,506,823	10,338,867
Syns+Cont.	RWS	0.0882	0.2692	37h 16m	23s 265ms	3,606,693	7,945,083
Ad hoc entity retrieval							
Lucene	TF-IDF	0.0373	0.0636	59h 17m	1s 370ms	–	–
	BM25	0.0668	0.1182		798ms	–	–
Fixed parameters over HGoE variations: $RWS(\ell = 2, r = 10^4, \Delta_{nf} = 0, \Delta_{ef} = 0, exp = F, wei = F)$							
Base Model	RWS	0.1390	0.2455	33h 53m	26s 330ms	3,506,823	7,721,743
Syns	RWS	0.1337	0.2473	33h 05m	30s 232ms	3,510,462	7,734,931
Cont.	RWS	0.1304	0.2364	34h 37m	27s 620ms	3,604,185	7,929,841
TF-Bins$_2$	RWS	0.0300	0.1145	35h 26m	4m 41s	3,506,823	10,338,867
Syns+Cont.	RWS	0.1313	0.2509	37h 16m	26s 877ms	3,606,693	7,945,083
Entity list completion							
Lucene	TF-IDF	0.0558	0.1000	59h 17m	1s 230ms	–	–
	BM25	0.0666	0.1250		1s 221ms	–	–
Fixed parameters over HGoE variations: $RWS(\ell = 2, r = 10^4, \Delta_{nf} = 0, \Delta_{ef} = 0, exp = F, wei = F)$							
Base Model	RWS	0.0879	0.0769	33h 53m	19s 162ms	3,506,823	7,721,743
Syns	RWS	0.0857	0.0635	33h 05m	19s 875ms	3,510,462	7,734,931
Cont.	RWS	0.0875	0.0692	34h 37m	19s 422ms	3,604,185	7,929,841
TF-Bins$_2$	RWS	0.0006	0.0058	35h 26m	1m 08s	3,506,823	10,338,867
Syns+Cont.	RWS	0.0884	0.0788	37h 16m	19s 824ms	3,606,693	7,945,083

Evaluation: INEX 2009 Wikipedia full collection

Multiple tasks

Keyword-based document profiles

Note: The entity index was created from sentence-based entity profiles

Table 10.6: Overall comparison of retrieval performance, for multiple entity-oriented search tasks, based on the complete INEX 2009 Wikipedia collection.

Index	Ranking	MAP	P@10	Index Time	Avg./Query	Nodes	Edges
Ad hoc document retrieval							
Lucene	TF-IDF	0.0228	0.0692	15h 06m	460ms	–	–
	BM25	0.0324	0.1173		370ms	–	–
Fixed parameters over HGoE variations: RWS($\ell = 2, r = 10^4, \Delta_{nf} = 0, \Delta_{ef} = 0, exp = F, wei = F$)							
Base Model	RWS	0.0863	0.2462	33h 53m	23s 405ms	3,506,823	7,721,743
Syns	RWS	0.0937	0.2615	33h 05m	55s 555ms	3,510,462	7,734,931
Cont.	RWS	0.0869	0.2654	34h 37m	24s 348ms	3,604,185	7,929,841
TF-Bins$_2$	RWS	0.0172	0.0500	35h 26m	2m 58s	3,506,823	10,338,867
Syns+Cont.	RWS	0.0882	0.2692	37h 16m	23s 265ms	3,606,693	7,945,083
Ad hoc entity retrieval							
Lucene	TF-IDF	0.0373	0.0636	59h 17m	1s 370ms	–	–
	BM25	0.0668	0.1182		798ms	–	–
Fixed parameters over HGoE variations: RWS($\ell = 2, r = 10^4, \Delta_{nf} = 0, \Delta_{ef} = 0, exp = F, wei = F$)							
Base Model	RWS	0.1390	0.2455	33h 53m	26s 330ms	3,506,823	7,721,743
Syns	RWS	0.1337	0.2473	33h 05m	30s 232ms	3,510,462	7,734,931
Cont.	RWS	0.1304	0.2364	34h 37m	27s 620ms	3,604,185	7,929,841
TF-Bins$_2$	RWS	0.0300	0.1145	35h 26m	4m 41s	3,506,823	10,338,867
Syns+Cont.	RWS	0.1313	0.2509	37h 16m	26s 877ms	3,606,693	7,945,083
Entity list completion							
Lucene	TF-IDF	0.0558	0.1000	59h 17m	1s 230ms	–	–
	BM25	0.0666	0.1250		1s 221ms	–	–
Fixed parameters over HGoE variations: RWS($\ell = 2, r = 10^4, \Delta_{nf} = 0, \Delta_{ef} = 0, exp = F, wei = F$)							
Base Model	RWS	0.0879	0.0769	33h 53m	19s 162ms	3,506,823	7,721,743
Syns	RWS	0.0857	0.0635	33h 05m	19s 875ms	3,510,462	7,734,931
Cont.	RWS	0.0875	0.0692	34h 37m	19s 422ms	3,604,185	7,929,841
TF-Bins$_2$	RWS	0.0006	0.0058	35h 26m	1m 08s	3,506,823	10,338,867
Syns+Cont.	RWS	0.0884	0.0788	37h 16m	19s 824ms	3,606,693	7,945,083

Evaluation: INEX 2009 Wikipedia full collection

Multiple tasks

Keyword-based document profiles

Note: The entity index was created from sentence-based entity profiles

Table 10.6: Overall comparison of retrieval performance, for multiple entity-oriented search tasks, based on the complete INEX 2009 Wikipedia collection.

Index	Ranking	MAP	P@10	Index Time	Avg./Query	Nodes	Edges
Ad hoc document retrieval							
Lucene	TF-IDF	0.0228	0.0692	15h 06m	460ms	–	–
	BM25	0.0324	0.1173		370ms	–	–
Fixed parameters over HGoE variations: $RWS(\ell = 2, r = 10^4, \Delta_{nf} = 0, \Delta_{ef} = 0, exp = F, wei = F)$							
Base Model	RWS	0.0863	0.2462	33h 53m	23s 405ms	3,506,823	7,721,743
Syns	RWS	0.0937	0.2615	33h 05m	55s 555ms	3,510,462	7,734,931
Cont.	RWS	0.0869	0.2654	34h 37m	24s 348ms	3,604,185	7,929,841
TF-Bins$_2$	RWS	0.0172	0.0500	35h 26m	2m 58s	3,506,823	10,338,867
Syns+Cont.	RWS	0.0882	0.2692	37h 16m	23s 265ms	3,606,693	7,945,083
Ad hoc entity retrieval							
Lucene	TF-IDF	0.0373	0.0636	59h 17m	1s 370ms	–	–
	BM25	0.0668	0.1182		798ms	–	–
Fixed parameters over HGoE variations: $RWS(\ell = 2, r = 10^4, \Delta_{nf} = 0, \Delta_{ef} = 0, exp = F, wei = F)$							
Base Model	RWS	0.1390	0.2455	33h 53m	26s 330ms	3,506,823	7,721,743
Syns	RWS	0.1337	0.2473	33h 05m	30s 232ms	3,510,462	7,734,931
Cont.	RWS	0.1304	0.2364	34h 37m	27s 620ms	3,604,185	7,929,841
TF-Bins$_2$	RWS	0.0300	0.1145	35h 26m	4m 41s	3,506,823	10,338,867
Syns+Cont.	RWS	0.1313	0.2509	37h 16m	26s 877ms	3,606,693	7,945,083
Entity list completion							
Lucene	TF-IDF	0.0558	0.1000	59h 17m	1s 230ms	–	–
	BM25	0.0666	0.1250		1s 221ms	–	–
Fixed parameters over HGoE variations: $RWS(\ell = 2, r = 10^4, \Delta_{nf} = 0, \Delta_{ef} = 0, exp = F, wei = F)$							
Base Model	RWS	0.0879	0.0769	33h 53m	19s 162ms	3,506,823	7,721,743
Syns	RWS	0.0857	0.0635	33h 05m	19s 875ms	3,510,462	7,734,931
Cont.	RWS	0.0875	0.0692	34h 37m	19s 422ms	3,604,185	7,929,841
TF-Bins$_2$	RWS	0.0006	0.0058	35h 26m	1m 08s	3,506,823	10,338,867
Syns+Cont.	RWS	0.0884	0.0788	37h 16m	19s 824ms	3,606,693	7,945,083

Evaluation: INEX 2009 Wikipedia full collection

Multiple tasks

Keyword-based document profiles

Note: The entity index was created from sentence-based entity profiles

Table 10.6: Overall comparison of retrieval performance, for multiple entity-oriented search tasks, based on the complete INEX 2009 Wikipedia collection.

Index	Ranking	MAP	P@10	Index Time	Avg./Query	Nodes	Edges
AD HOC DOCUMENT RETRIEVAL							
Lucene	TF-IDF	0.0228	0.0692	15h 06m	460ms	–	–
	BM25	0.0324	0.1173		370ms	–	–
Fixed parameters over HGoE variations: RWS($\ell = 2, r = 10^4, \Delta_{nf} = 0, \Delta_{ef} = 0, exp = F, wei = F$)							
Base Model	RWS	0.0863	0.2462	33h 53m	23s 405ms	3,506,823	7,721,743
Syns	RWS	0.0937	0.2615	33h 05m	55s 555ms	3,510,462	7,734,931
Cont.	RWS	0.0869	0.2654	34h 37m	24s 348ms	3,604,185	7,929,841
TF-Bins$_2$	RWS	0.0172	0.0500	35h 26m	2m 58s	3,506,823	10,338,867
Syns+Cont.	RWS	0.0882	0.2692	37h 16m	23s 265ms	3,606,693	7,945,083
AD HOC ENTITY RETRIEVAL							
Lucene	TF-IDF	0.0373	0.0636	59h 17m	1s 370ms	–	–
	BM25	0.0668	0.1182		798ms	–	–
Fixed parameters over HGoE variations: RWS($\ell = 2, r = 10^4, \Delta_{nf} = 0, \Delta_{ef} = 0, exp = F, wei = F$)							
Base Model	RWS	0.1390	0.2455	33h 53m	26s 330ms	3,506,823	7,721,743
Syns	RWS	0.1337	0.2473	33h 05m	30s 232ms	3,510,462	7,734,931
Cont.	RWS	0.1304	0.2364	34h 37m	27s 620ms	3,604,185	7,929,841
TF-Bins$_2$	RWS	0.0300	0.1145	35h 26m	4m 41s	3,506,823	10,338,867
Syns+Cont.	RWS	0.1313	0.2509	37h 16m	26s 877ms	3,606,693	7,945,083
ENTITY LIST COMPLETION							
Lucene	TF-IDF	0.0558	0.1000	59h 17m	1s 230ms	–	–
	BM25	0.0666	0.1250		1s 221ms	–	–
Fixed parameters over HGoE variations: RWS($\ell = 2, r = 10^4, \Delta_{nf} = 0, \Delta_{ef} = 0, exp = F, wei = F$)							
Base Model	RWS	0.0879	0.0769	33h 53m	19s 162ms	3,506,823	7,721,743
Syns	RWS	0.0857	0.0635	33h 05m	19s 875ms	3,510,462	7,734,931
Cont.	RWS	0.0875	0.0692	34h 37m	19s 422ms	3,604,185	7,929,841
TF-Bins$_2$	RWS	0.0006	0.0058	35h 26m	1m 08s	3,506,823	10,338,867
Syns+Cont.	RWS	0.0884	0.0788	37h 16m	19s 824ms	3,606,693	7,945,083

IV

Conclusions

Discussion, final remarks and future work

Discussion

Efficiency / effectiveness trade-off:

- Lower r is more efficient
- Higher r is more effective

Current implementation is:

- Less efficient when compared to Lucene
- More effective in the experiments using keyword-based document profiles

Discussion

Ad hoc document retrieval

- Best MAP: **0.1689** (vs 0.1710 TF-IDF)
  - Base model
  - INEX 2009 10T-NL
  - ℓ=2, r=10,000, exp.=true
- Best P@10: **0.2692** (vs 0.0692 TF-IDF)
  - Synonyms+Context model
  - INEX 2009 full collection
  - ℓ=2, r=10,000, exp.=false

Note: In bold the best scores for the hypergraph-of-entity; in parenthesis the baseline result of Lucene TF-IDF for the same experiment.

Discussion

Ad hoc entity retrieval

- Best MAP: **0.1390** (vs 0.0373 TF-IDF)
  - Base model
  - INEX 2009 full collection
  - ℓ=2, r=10,000, exp.=false
- Best P@10: **0.2509** (vs 0.0636 TF-IDF)
  - Synonyms+Context model
  - INEX 2009 full collection
  - ℓ=2, r=10,000, exp.=false

Entity list completion

- Best MAP: **0.0884** (vs 0.0558 TF-IDF)
  - Synonyms+Context model
  - INEX 2009 full collection
  - ℓ=2, r=10,000, exp.=false
- Best P@10: **0.0788** (vs 0.1000 TF-IDF)
  - Synonyms+Context model
  - INEX 2009 full collection
  - ℓ=2, r=10,000, exp.=false

Note: In bold the best scores for the hypergraph-of-entity; in parenthesis the baseline result of Lucene TF-IDF for the same experiment.

Final remarks

- We have proven that a graph-based model is viable in EOS...
- ...as a joint representation of corpora and knowledge bases...
- ...using a universal ranking function to solve multiple EOS tasks.

- We improved retrieval effectiveness in some particular cases...
- ...motivating the continued research of hypergraph-based models...
- ...and unified frameworks in information retrieval.

Future work

Improve random walk score efficiency:

- Algebraic approach

Improve the quality of the hypergraph:

- Prune nodes and hyperedges
- Reduce the cost of experiments to explore different types of relations

Further explore and improve:

- Node and hyperedge weighting functions (TF-bins in particular)

Expand the model to other tasks, e.g.:

- Personalized search (user profiles)
- Text augmentation (with entities)
- Lexicon construction (term ranking)

THANK YOU!

Open to discussion

You can find Army ANT's code at (check 'devel' branch):
https://github.com/feup-infolab/army-ant

And visit the ANT search engine at:
https://ant.fe.up.pt/

Appendix

Extra detail to aid discussion

Consolidating models

- From physics to machine learning
  - Efforts to unify theories and models
- Towards general approaches to IR
  - Identifying commonalities along the pipeline and tasks
- Graphs as general representation models
  - Combining all available information sources
- Unified framework for IR
  - Solving the information need is the only task

Classical models

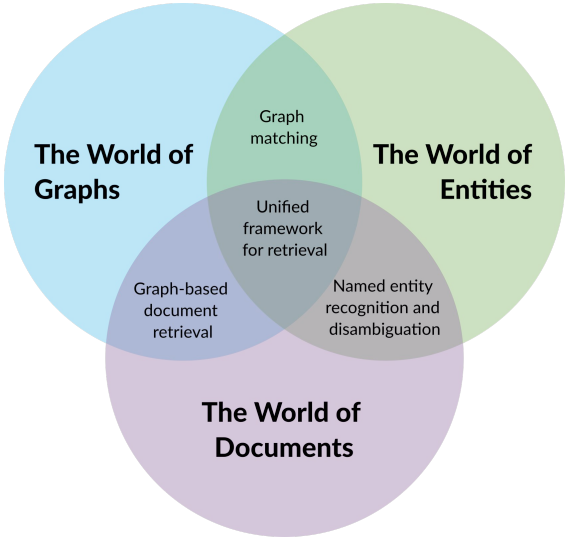
- Ranking
  - From TF-IDF
  - To Markov networks
- Representation
  - Virtual documents
  - Triplestores
- Hybrid approaches
  - Combined signals (single task or chained tasks)
  - Joint indexing of text and triples (very few contributions)

Learning-to-rank models

- Entity ranking based on:
  - Entity features
    - Text features from the Wikipedia article
    - Graph features from the knowledge graph
  - Entity profiles represented as virtual documents
    - “Flattened” data from RDF
    - Passages of text mentioning the entity
- Joint learning of representations for words and entities

Graph-based models

- Link analysis
- Text as a graph
- Knowledge graphs
- Entity graph from text
- Entity graph as a tensor
- Graph matching
- Hypergraph-based
- Random walk based



Anchor / core references

Title:
Entity-Oriented Search

Authors:
K. Balog

Year:
2018

DOI:
10.1007/978-3-319-93935-3

- First complete reference in the area
- Clear definitions of fundamental concepts
- Identification of tasks and applications
- Provides a compilation and a convergence of information

Anchor / core references

Title:
Concordance-Based Entity-Oriented Search

Authors:
M. Bautin and S. Skiena

Year:
2007

DOI:
10.1109/WI.2007.84

- Analyzes the presence of entities in queries (based on AOL query log):
  - 18-39% queries directly reference entities
  - 73-87% queries contain at least one entity
- "First-in-literature" implementation of an entity search engine
- Archetype for approaches based on virtual documents

Anchor / core references

Title:
Ad-hoc object retrieval in the web of data

Authors:
J. Pound, P. Mika, and H. Zaragoza

Year:
2010

DOI:
10.1145/1772690.1772769

Five query categories for ad hoc entity retrieval:

- Entity query
- Type query
- Attribute query
- Relation query
- Keyword query

Applied to the ANT search engine

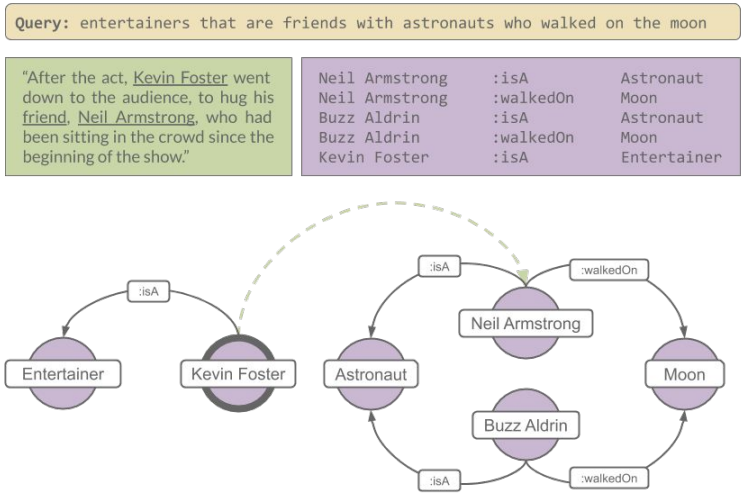
Anchor / core references

Title:
An Index for Efficient Semantic Full-text Search

Authors:
H. Bast and B. Buchhold

Year:
2013

DOI:
10.1145/2505515.2505689



Anchor / core references

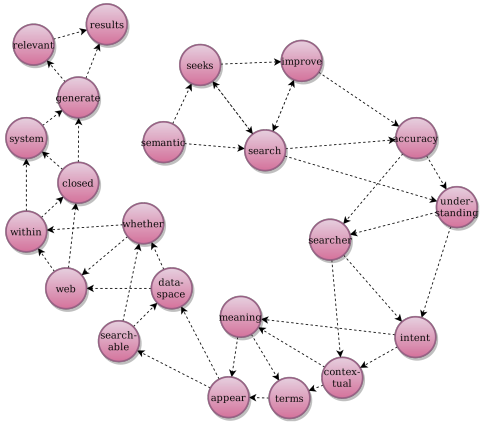
Title:
Graph-of-word and TW-IDF: new approach to ad hoc IR

Authors:
F. Rousseau and M. Vazirgiannis

Year:
2013

DOI:
10.1145/2505515.2505671

- Nodes represent terms
- Edges represent following terms within a window of size N
- TW is given by the indegree



Anchor / core references

Title:
Modeling Higher-order Term Dependencies in Information Retrieval Using Query Hypergraphs

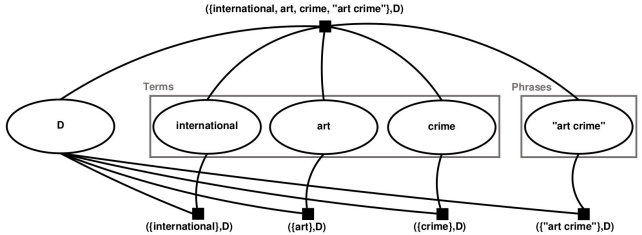
Authors:
M. Bendersky and W. B. Croft

Year:
2012

DOI:
10.1145/2348283.2348408

Query hypergraph model

- Log-linear retrieval model
- Solved through a factor graph
- Similar to Markov networks
- But captures higher-order relations (e.g., bigrams, named entities)



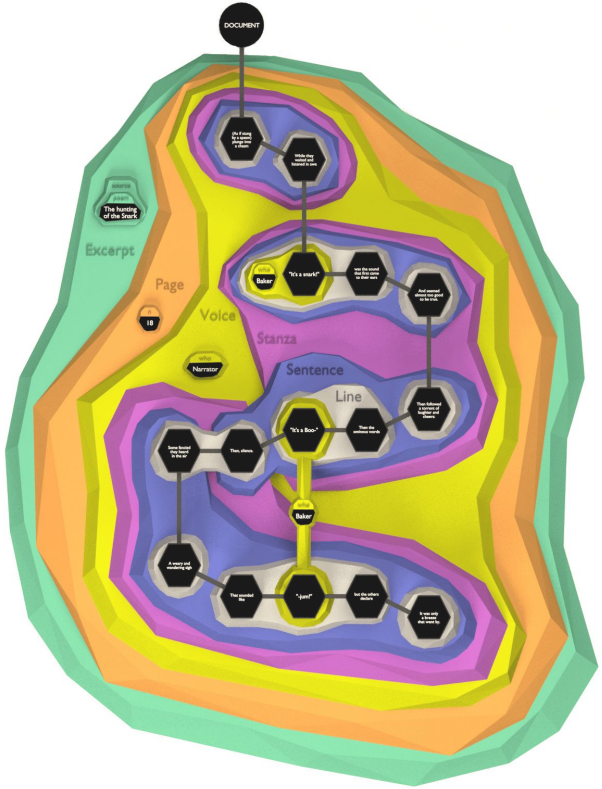
Anchor / core references

Title:
It's more than just overlap: Text As Graph

Authors:
R. Haentjens Dekker and D. J. Birnbaum

Year:
2017

DOI:
10.4242/BalisageVol19.Dekker01



COMBINED DATA

Combined data is a collection of corpora and knowledge bases, which includes not only the natural relations between documents (e.g., hyperlinks in the web), and entities (e.g., object properties in triplestores), but also cross-context relations, from mentions found in documents to entities in knowledge bases, and from entities found in knowledge bases to instances of the same entity in other knowledge bases.

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.

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.

Contributed datasets

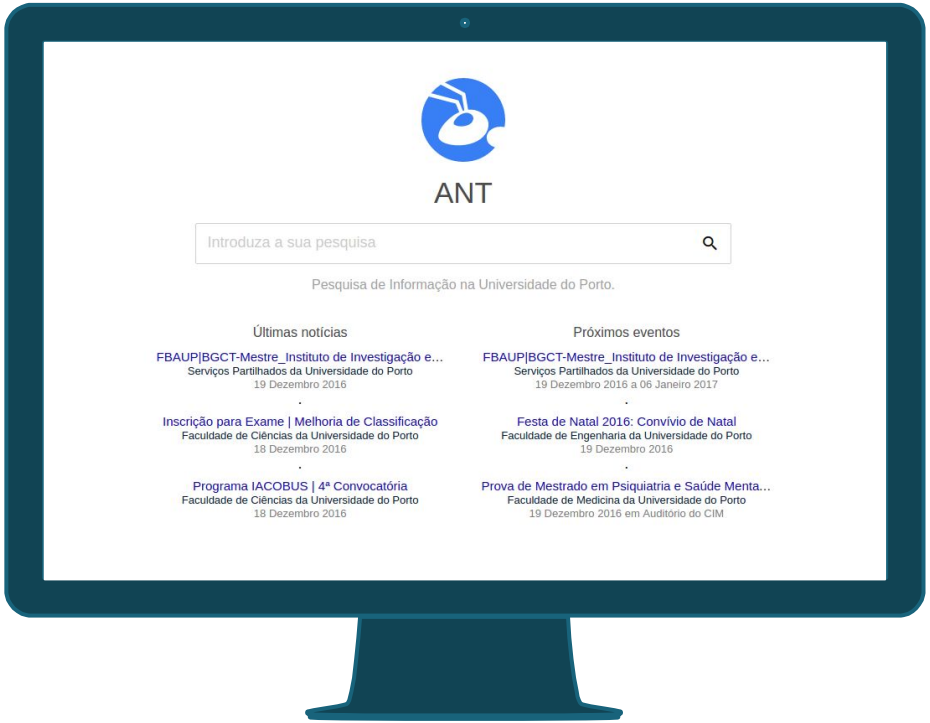
Simple English Wikipedia Link Graph with Clickstream Transitions 2018-12

DOI: 10.25747/83vk-zt74

ANT

Entity-oriented search engine for the University of Porto

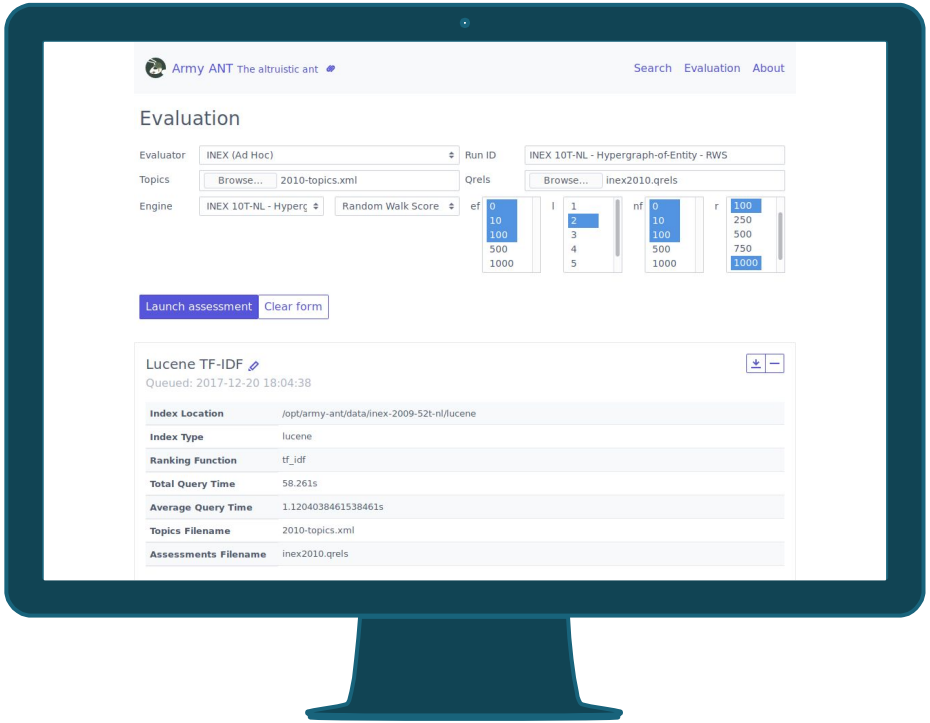
- Working prototype (https://ant.fe.up.pt)
- Exposure to ~1,000 weekly users
- Manifested interest by some of the faculty's content managers



Army ANT

Workbench for innovation in entity-oriented search

- Promotes freedom and exploration
- Supports IR research in a flexible way
- Available at GitHub
  (https://github.com/feup-infolab/army-ant)



Hypergraph-of-entity: Characterization

Basic statistics over time (i.e., as the index grows):

- Number of nodes and hyperedges per type and direction
- Node-based and hyperedge-based degree distributions
- Hyperedge cardinality distribution
- Clustering coefficient
- Average path length & diameter
- Density
- Shortest distances computed based on random walks
- Two-node clustering coefficients
  - Based on a set of sampled nodes
  - And a large sample of their neighbors
- Density based on a corresponding bipartite graph

$$D = \frac{2\sum_k k|E_U^k| + \sum_{k_1,k_2}(k_1+k_2)|E_D^{k_1,k_2}|}{2(n+m)(n+m-1)}$$

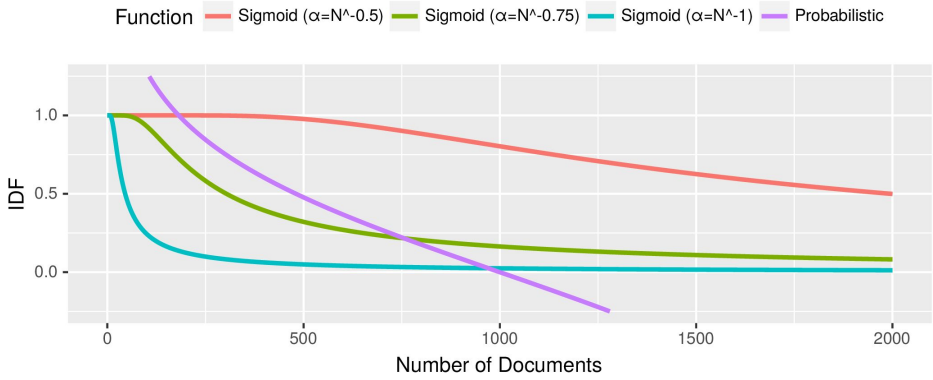
Hypergraph-of-entity: Joint representation model

Extensions: weights

Table 7.2: Hypergraph-of-entity weighting functions.

(a) Nodes.

Node / Weight	Description
term $2S\left(\alpha\frac{N-n_t}{n_t}\right)-1$	We used a variation of the IDF, with a tunable $\alpha < 1$ parameter to control how fast the function decreases. - S is the sigmoid function - N is the number of documents in the collection - n_t is the number of documents where a given term t occurs. - We used $\alpha = N^{-0.75}$.
entity Same as *term*.	In the future, we will experiment with different values of α for terms and entities, in particular alternative exponents to -0.75.



Hypergraph-of-entity: Joint representation model

Extensions: weights

(b) Hyperedges.

Hyperedge / Weight	Description
document 0.5	Linking a term or entity simply through document co-occurrence is weak, so we use a constant weight lower than one.
related_to $\frac{1}{\lvert e_r \rvert} \sum_{v \in e_r} \frac{\lvert \{u \in e'_r : e'_r \in E_r \setminus \{e_r\} \wedge v \in e'_r\} \rvert}{\lvert e_r \rvert}$	For each entity within the hyperedge, we calculate the fraction of reachable other entities and average all results. - E_r is the set of all *related_to* hyperedges. - $e_r \in E_r$ is the specific *related_to* hyperedge, for which we are calculating the weight.
contained_in $\frac{1}{\lvert t \rvert}$	Links with fewer terms t, where t refers to the tail set in $(t, h) \in E_c \wedge t \subseteq V_t$, should be more frequently followed, since the certainty that the hyperedge leads to the entity is higher.
synonym $\frac{1}{\lvert e_s \rvert}$	The higher the number of possible synonyms $e_s \in E_s \wedge e_s \subseteq V_t$, the less certain we are about the hyperedge — we rely on the synonyms of the first (and most probable) sense according to WordNet.
context $\frac{1}{\lvert e_x \rvert} \sum_{t_i \in e_x \setminus \{t_k\}} \frac{\mathrm{sim}(t_k, t_i) - \mathrm{min}_{\mathrm{sim}}}{1 - \mathrm{min}_{\mathrm{sim}}}$	A context $e_x \in E_x$ is only as good as the average of all similarities between the original term $t_k \in e_x$ and all other terms $t_i \in e_x \setminus \{t_k\}$. We normalize the weight taking into account the threshold used to create the word2vec SimNet.

Evaluation: Main experiments

Rank stability:

- Average Kendall's coefficient of concordance
  - Over 100 similar runs per configuration
  - For different values of ℓ and r
- For $\ell \in \{2, 3, 4\}$:
  - 84–90% stable for r=100
  - 94–97% stable for r=1,000
  - 99% stable for r=10,000

Evaluation: INEX 2009 Wikipedia collection

Three subsets:

- INEX 2009 3T-NL (2.2k docs)
- INEX 2009 10T-NL (7.5k docs)
- INEX 2009 52T-NL (37.8k docs)

Created through:

- Random sampling of n topics
- Retained relevance judgments for selected topics
- Retained only judged documents

However:

- Exclusively for assessing ad hoc document retrieval
- Based on the qrels for the 2010 Ad Hoc track
- Goal: eventually index the full collection
- Challenge: scalability