

KNOWLEDGE GRAPH-BASED SYSTEM FOR TECHNICAL DOCUMENT RETRIEVAL

A DEDUCTIVE REASONING-FOCUSED EXPLORATION

PhD student: Matthias SESBOÛÉ

Directed by Cecilia ZANNI-MERK

Supervised by Nicolas DELESTRE and Jean-Philippe
KOTOWICZ

September 5, 2024



RESPONDING THESIS

Knowledge Graph-based System for Technical Document Retrieval

A deductive reasoning-focused exploration

- Research objective: Leveraging domain knowledge to enhance Information Retrieval in a technical context.
- Traceparts employment partially financed by ANRT as part of a CIFRE research project in collaboration with the Litis lab.
- Began on March 15th 2021.

TRACEPARTS

One of the world's leading CAD-content platforms for Engineering, Industrial Equipment and Machine Design. The CAD-content platform *traceparts.com* provides access to over 1.8 thousand supplier-certified product catalogues with 2D drawings, 3D CAD models and product datasheets.

- Technical content aimed at an engineering audience from multiple industries
- Content available in 25 languages
- Users can search using :
 - A full text search
 - A list of catalogues
 - Different classifications



CORPUS

- Over 1.1 million document families
- Over 127.8 millions individual documents
- 25 languages
- Documents' texts contain average 50 characters and 7 words
- Over 210 thousand tags, amongst which:
 - Over 2.5 thousand suppliers and manufacturers
 - Over 1.9 thousand catalogues
 - Over 208 thousand categories

Some text content examples are:

- *DIN 912*
- *The P01 to P08 pumps are designed to pump lubricating fluids (oil, diesel oil, etc.). Their flow rate is from 1 to 24 L / min; maximum working pressure 10 bar.*

USER SEARCHES

User text searches:

- Are composed of domain-specific keywords, notations, identifiers, and acronyms.
- Contain on average 13 characters separated into 2 words.
- Can come in any language

The most common searches are:

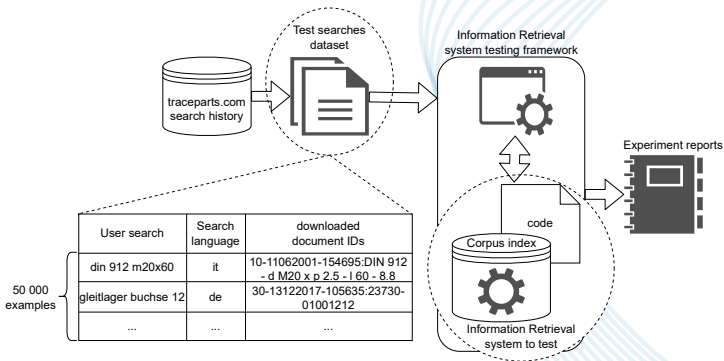
- *motor*
- *din 912*
- *ball valve*

TRACEPARTS SEARCH SYSTEM CHALLENGES

Traceparts search challenges come from:

- Short multilingual texts
- Technical texts with many synonyms, acronyms, homonyms, and notations
- A large and heterogeneous corpus
- Multiple engineering domains coverage
- High recall but low precision

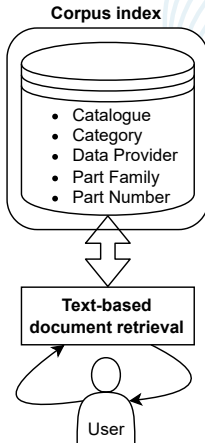
EXPERIMENTAL PROTOCOL



EVALUATION METRICS

- Mean Average Precision at k (MAP@k):
 - A sliding (or growing) precision window, averaged over a set of query examples.
 - Ranges from 0 to 1 (1 is the best value).
 - Gives information about the amount and positions of positive results in the k first ones.
- Binary Mean at k (BM@k):
 - Binary average over a set of query examples.
 - Ranges from 0 to 1 (1 is the best value).
 - Provides information about the amount of queries with a positive result in the k first ones.
 - Does not give any detail on the positive result position.

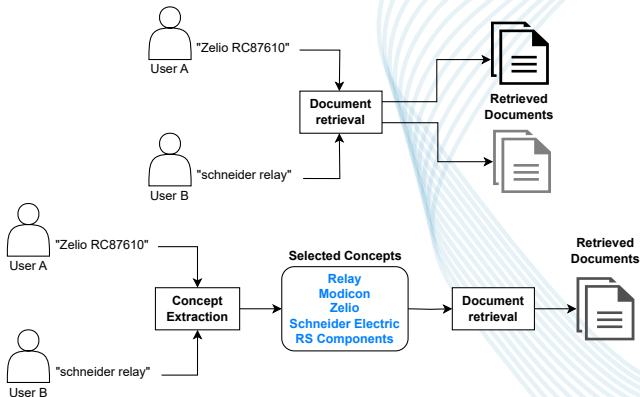
TEXT-BASED SYSTEM (BASELINE)



TEXT-BASED SYSTEM (BASELINE) RESULTS

Text-based system (baseline)		
@k ↓	MAP@k	BM@k
@5	0.061	0.114
@25	0.064	0.148
@50	0.064	0.157
@100	0.064	0.161
@350	0.064	0.164

OUR APPROACH BASED ON A KNOWLEDGE GRAPH



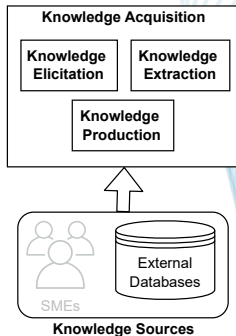
KNOWLEDGE GRAPH AND ONTOLOGY

Knowledge Graph (Hogan et. al. 2021): *a knowledge graph is a graph of data intended to accumulate and convey knowledge of the real world, whose nodes represent entities of interest and whose edges represent relations between these entities.*

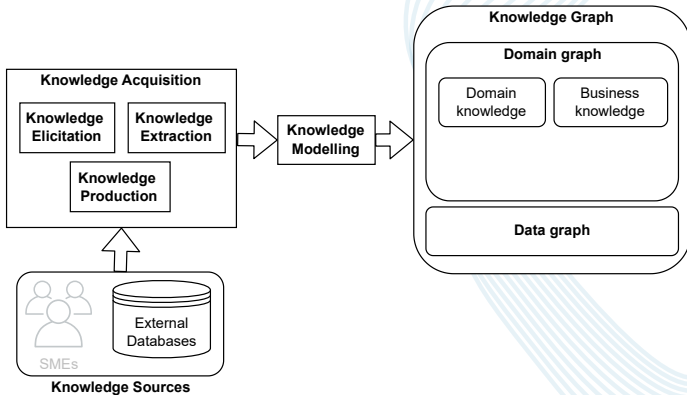
Ontology (Hogan et. al. 2021): *In the context of computing, an ontology is then a concrete, formal representation of what terms mean within the scope in which they are used (e.g., a given domain).*

In our work, we consider an ontology a particular component of
a Knowledge Graph

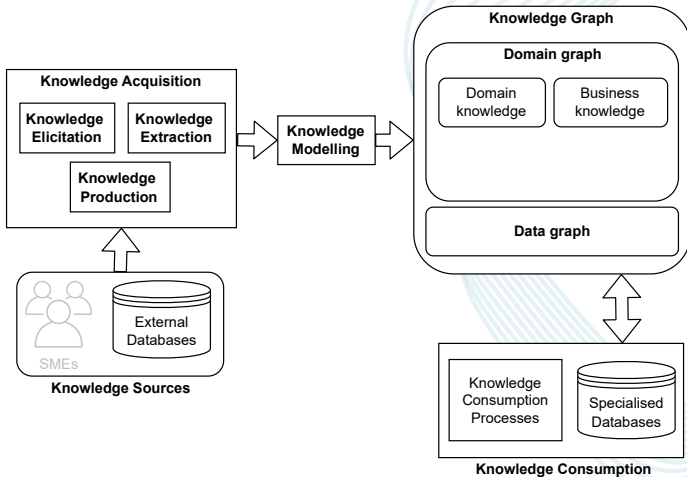
KNOWLEDGE ACQUISITION



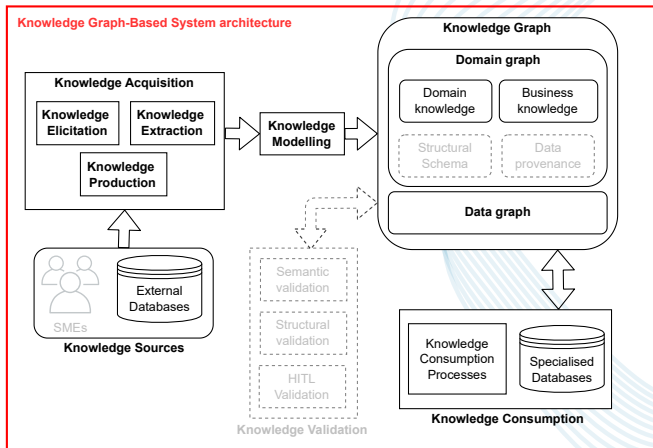
KNOWLEDGE GRAPH



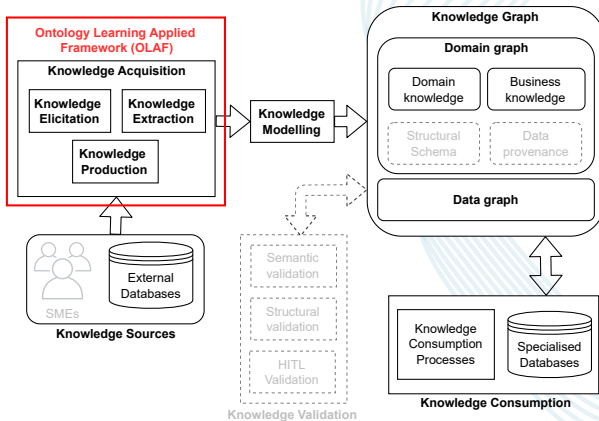
KNOWLEDGE CONSUMPTION



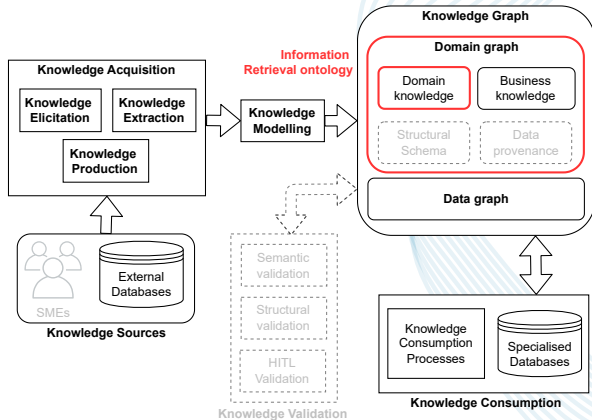
KNOWLEDGE GRAPH-BASED SYSTEM ARCHITECTURE



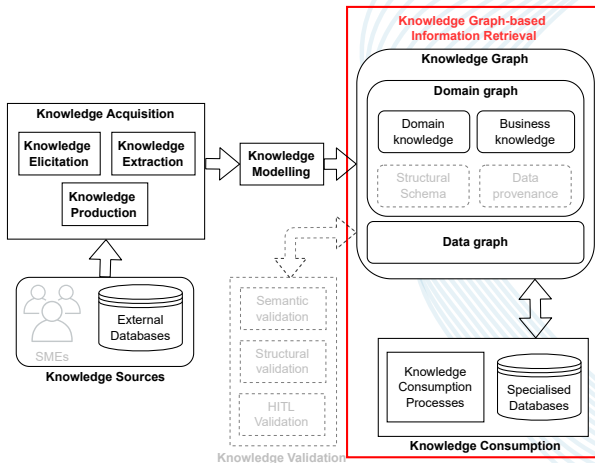
ONTOLOGY LEARNING APPLIED FRAMEWORK (OLAF)



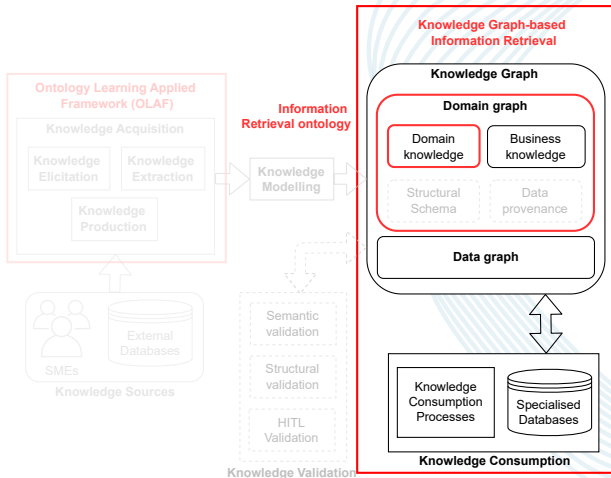
INFORMATION RETRIEVAL ONTOLOGY



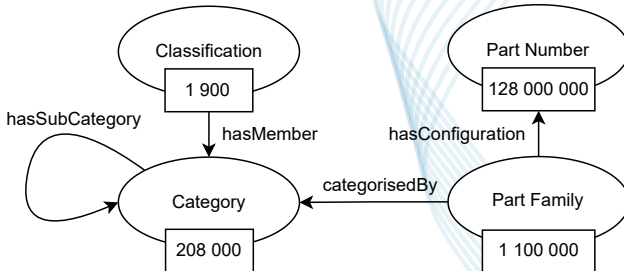
KG-BASED INFORMATION RETRIEVAL



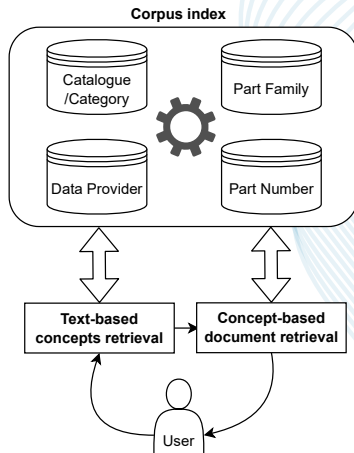
IN THIS PRESENTATION



TRACEPARTS KNOWLEDGE GRAPH



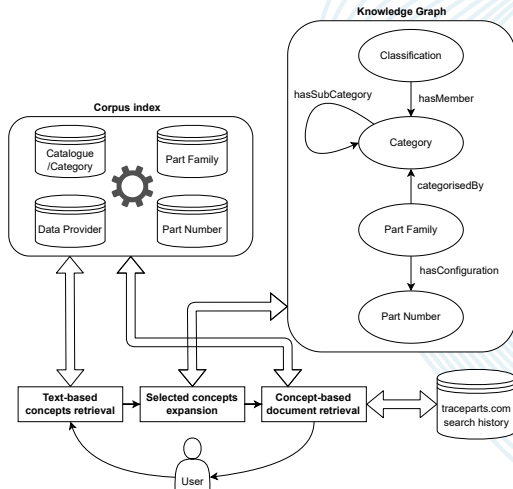
CONCEPT-BASED SYSTEM



CONCEPT-BASED SYSTEM RESULTS

@k ↓	Text-based system (baseline)		Concept-based system	
	MAP@k	BM@k	MAP@k	BM@k
@5	0.061	0.114	0.152	0.243
@25	0.064	0.148	0.159	0.334
@50	0.064	0.157	0.160	0.371
@100	0.064	0.161	0.161	0.403
@350	0.064	0.164	0.161	0.429

KG-BASED SYSTEM WITH IMPLICIT KNOWLEDGE



KG-BASED SYSTEM WITH IMPLICIT KNOWLEDGE RESULTS

	Text-based system (baseline)		Concept-based system		KG-based system with search history	
@k ↓	MAP@k	BM@k	MAP@k	BM@k	MAP@k	BM@k
@5	0.061	0.114	0.152	0.243	0.115	0.291
@25	0.064	0.148	0.159	0.334	0.122	0.471
@50	0.064	0.157	0.160	0.371	0.123	0.552
@100	0.064	0.161	0.161	0.403	0.123	0.624
@350	0.064	0.164	0.161	0.429	0.124	0.715

ONLINE OWL REASONING-BASED APPROACH

An approach focusing on OWL.

- An Information Retrieval ontology.
- Push knowledge closer to the data.
- Model domain knowledge as linked sets of taxonomies.

Competency questions:

- CQ1 What are the categories in the user search?
- CQ2 What are the documents relevant to a search?
- CQ3 What categories are enabled to refine the search?

INFORMATION RETRIEVAL ONTOLOGY

7 classes:

- *CandidateDocument* subclass of *Document*
- *SelectedCategory* and *EnabledCategory* subclasses of *Category*
- *SearchContext* subclass of *Search*

6 object properties:

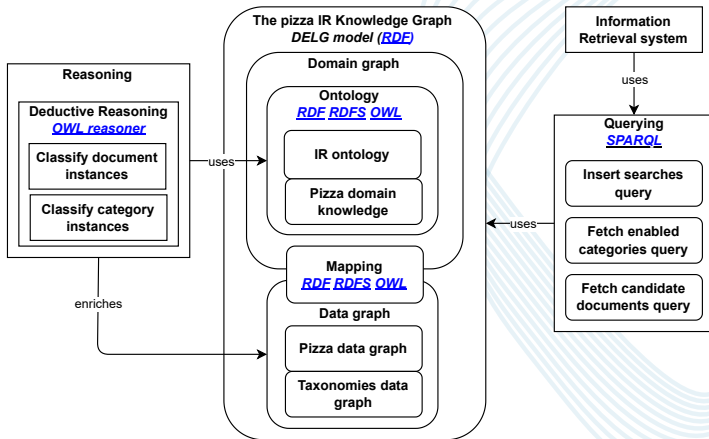
- *categorises* inverse of *categorisedBy*
- *hasSearchCategory* subproperty of *enablesCategory*
- *hasDirectSubcategory* subproperty of *hasSubcategory*

PIZZA ONTOLOGY

Pizza ontology:

- Well-knowledge ontology built to introduce RDF/RDFS/OWL with examples (and even SHACL)
- Simple ontology with class hierarchies of:
 - *Pizza* (*hasTopping*, *hasBase*)
 - *PizzaBase*
 - *PizzaTopping*
- We use the Pizza ontology for demonstration in interest of time constraints

PIZZA ONTOLOGY KNOWLEDGE GRAPH



CONCLUSION

We have:

- Explored a Knowledge Graph-Based System (KGBS) architecture
- Detailed each KGBS containers and activities
- Explored a real-world use case moving from a text-based to a KG-based Information Retrieval (IR) System.
- Introduce and compared 3 IR systems:
 - A text-based IR system
 - A concept-based IR system
 - A KG-based IR system
- Presented our Information Retrieval ontology

KG-based systems for IR on a multilingual corpus of technical documents show promising results overcoming the text-based approaches limitations.

FUTURE WORKS

- KGBS architecture:
 - Implement an end-to-end Knowledge Graph-Based System architecture use case.
 - Further explore the modularity of the architecture
- Knowledge Graph-based Information Retrieval system:
 - Expand the Knowledge Graph
 - Enhance the concept matching task
 - Expand the approach to other domains

CONTRIBUTIONS

Main contributions:

- Ontology Learning Applied Framework (OLAF)
- Knowledge Graph-based Information Retrieval systems
- An OWL Information Retrieval ontology

Satellite contributions:

- A unifying definition of Knowledge Graph
- An architecture for Knowledge Graph-Based Systems

SCIENTIFIC PRODUCTIONS

Peer-reviewed international conference papers:

- An operational architecture for knowledge graph-based systems. Proceedings of the 26th International Conference KES2022.
- (with Marion Schaeffer) Olaf: An ontology learning applied framework. Proceedings of the 27th International Conference KES2023.

Open-source software library (with Marion Schaeffer):

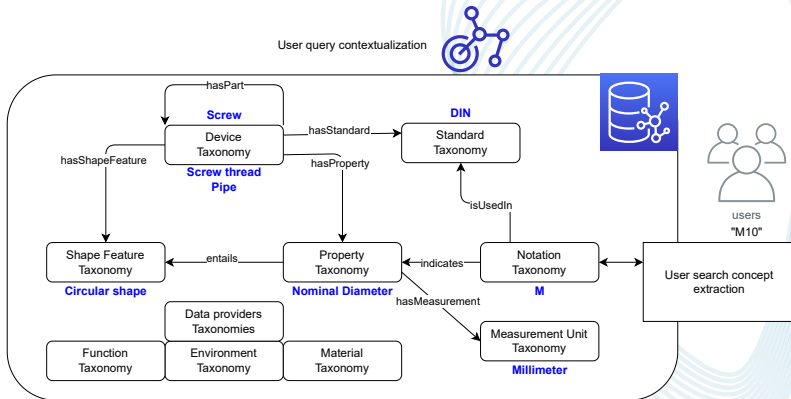
- Ontology Learning Applied Framework Python library implementation:
<https://wikit-ai.github.io/olaf/>

THANK YOU!

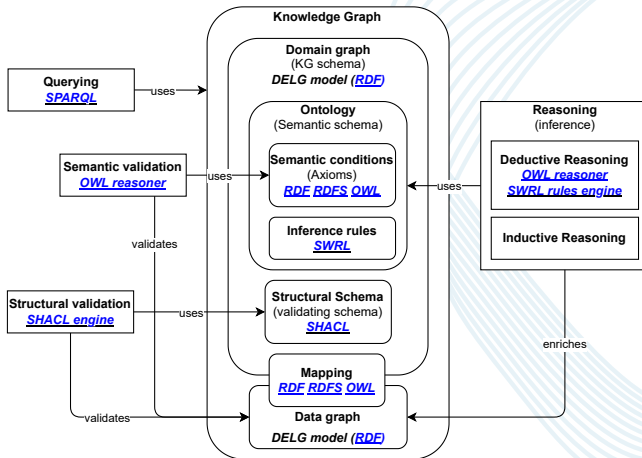
Thank you for your time and attention.

I am now ready to answer to any questions.

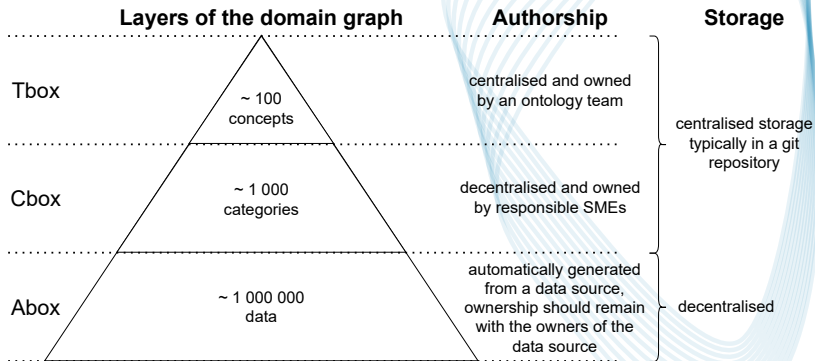
PERSPECTIVES: KNOWLEDGE GRAPH



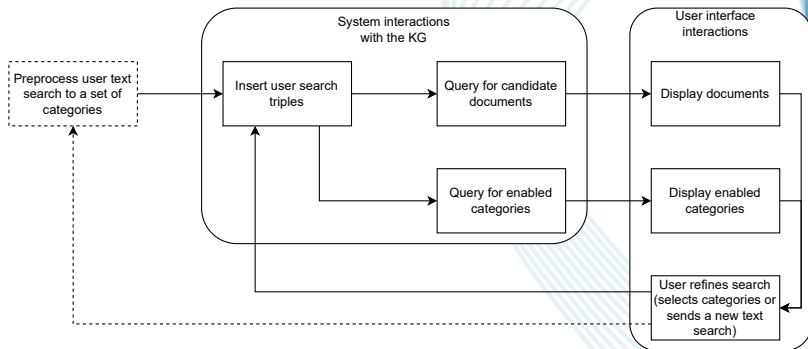
SEMANTIC WEB KNOWLEDGE GRAPH



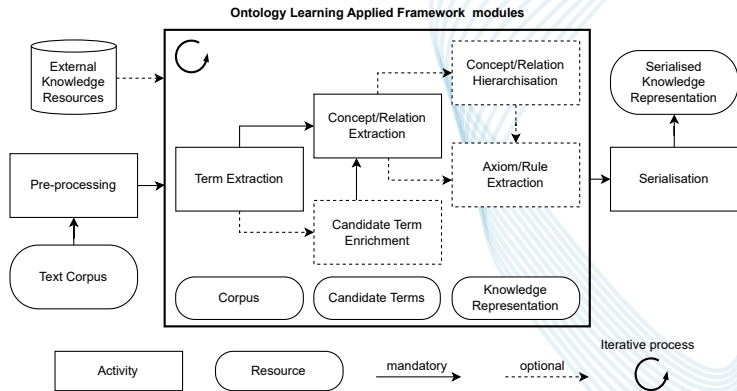
KNOWLEDGE MODELLING



OWL REASONING-BASED INFORMATION RETRIEVAL



ONTOLOGY LEARNING APPLIED FRAMEWORK



KG-BASED IR: QUANTITATIVE RESULTS

	No results	Less than 400 results (non empty)
Text-based system (baseline)	64.48%	35.44%
Concept-based system	11.43%	88.36%
KG-based system with search history	8.10%	51.59%