

Semantically enhanced Information Retrieval: an ontology-based approach

Miriam Fernández Sánchez

under the supervision of

Pablo Castells Azpilicueta

Departamento de Ingeniería Informática

Escuela Politécnica Superior

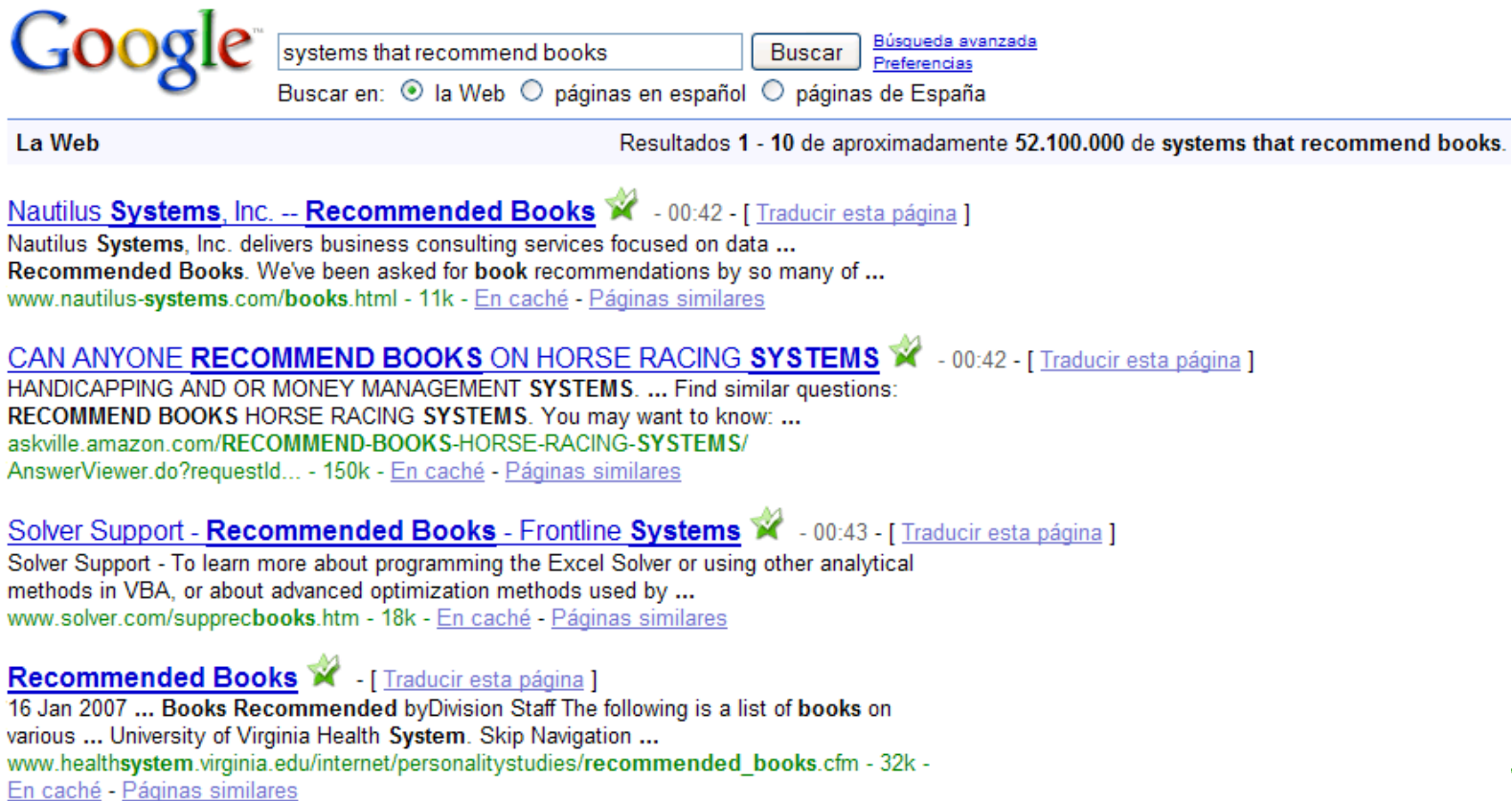
Universidad Autónoma de Madrid

Table of contents

- Motivation
- Part I. Analyzing the state of the art
 - What is semantic search?
- Part II. The proposal
 - An ontology-based IR model
 - Semantic retrieval on the Web
- Part III. Extensions
 - Semantic knowledge gateway
 - Coping with knowledge incompleteness
- Conclusions

Motivation (I)

- How to find and manage massive-scale content stored and shared on the Web and other large document repositories?
 - Using **search engines** (Google, Yahoo, MSN)
 - Do users always manage to find the information they are looking for?
 - **Example:** “systems that recommend books”



The screenshot shows a Google search interface with the query 'systems that recommend books'. The search results are displayed under the heading 'La Web' and show 'Resultados 1 - 10 de aproximadamente 52.100.000 de systems that recommend books.' The first three results are:

- Nautilus Systems, Inc. -- Recommended Books** - 00:42 - [Traducir esta página]
Nautilus Systems, Inc. delivers business consulting services focused on data ...
Recommended Books. We've been asked for **book** recommendations by so many of ...
www.nautilus-systems.com/books.html - 11k - [En caché](#) - [Páginas similares](#)
- CAN ANYONE RECOMMEND BOOKS ON HORSE RACING SYSTEMS** - 00:42 - [Traducir esta página]
HANDICAPPING AND OR MONEY MANAGEMENT **SYSTEMS**. ... Find similar questions:
RECOMMEND BOOKS HORSE RACING SYSTEMS. You may want to know: ...
askville.amazon.com/RECOMMEND-BOOKS-HORSE-RACING-SYSTEMS/AnswerViewer.do?requestId... - 150k - [En caché](#) - [Páginas similares](#)
- Solver Support - Recommended Books - Frontline Systems** - 00:43 - [Traducir esta página]
Solver Support - To learn more about programming the Excel Solver or using other analytical methods in VBA, or about advanced optimization methods used by ...
www.solver.com/supprecbooks.htm - 18k - [En caché](#) - [Páginas similares](#)

The fourth result is partially visible:

- Recommended Books** - [Traducir esta página]
16 Jan 2007 ... **Books Recommended** by Division Staff The following is a list of **books** on various ... University of Virginia Health **System**. Skip Navigation ...
www.healthsystem.virginia.edu/internet/personalitystudies/recommended_books.cfm - 32k - [En caché](#) - [Páginas similares](#)

Motivation (II)

- **Problem:** current content description and query processing techniques for IR are based on keywords
 - Limited capabilities to grasp and exploit the conceptualizations involved in user needs and content meanings
 - Relations between search terms: “books about recommender systems” vs. “systems that recommend books”
 - Polysemy: “jaguar” as animal vs. “jaguar” as car
 - Synonymy: “movies” vs. “films”
 - Documents about individuals where query keywords do not appear: “English banks”, individual “Abbey”
 - Inference: water sports in Mediterranean coast → windsurf, scuba diving, etc., in Valencia, Alicante, etc.
- **Potential solution:** semantic search
 - Search by meanings rather than literal strings
 - If machines are able to understand the semantics behind user needs and contents, they will retrieve more accurate results!



Motivation (III)

- **Semantic search:** different perspectives
 - Information Retrieval (IR)
 - The level of conceptualization is often shallow and sparse, especially at the level of relations
 - Semantic-based knowledge technologies (SW)
 - Use of higher levels of conceptualization but are more focused on data retrieval and not directly applied to unstructured information objects carrying free text or multimedia content
- **Goal:** The realization of a novel semantic retrieval model
 - Exploit deep levels of conceptualization
 - Support search in large, open and heterogeneous repositories of unstructured information

Research questions

- Q1: What do we understand by semantic search?
- Q2: Where are we standing in the progress towards semantic information retrieval?
- Q3: Can the achievements in semantic retrieval from different research fields be combined and give rise to enhanced retrieval models thereupon?
- Q4: Can semantic retrieval models be scaled to open, massive, heterogeneous environments such as the World Wide Web?
- Q5: How to standardize the evaluation of semantic retrieval systems?
- Q6: How to deal with knowledge incompleteness?

Part I. Analyzing the state of the art

- What is semantic search?
 - IR vs. semantic-based knowledge technologies perspectives
 - A global classification
 - Drawbacks and limitations

What is semantic search?

- What is semantic search?
 - Raising the representation of content meanings to a higher level above plain keywords, in order to enhance current mainstream Information Retrieval (IR) technologies
- Goal
 - Reduce the distance between the logic representation of the IR systems and the real one in the user's mind with regards to the formulation of queries and the understanding of documents
- Barriers
 - The bag of words approach is pervasively adopted in currently deployed IR technologies
 - Conceptual representations are difficult and costly to create and maintain

Semantic search: IR and SW perspectives

- Semantic search: IR perspective
 - Early 80s: elaboration of conceptual frameworks and their introduction in IR models
 - Taxonomies (categories + hierarchical relations) , e.g., Linaen taxonomy
 - Thesaurus (categories + fixed hierarchical & associative relations), e.g., WordNet (used by *linguistic approaches*)
 - Algebraic methods such as *LSA*
 - Main limitations: The level of conceptualization is often shallow (relations)
- Semantic search: semantic-based technologies perspective
 - Late 90s: introduction of **ontologies** as conceptual framework (classes + instances (KBs) + arbitrary semantic relations + rules)
 - Main limitations:
 - Semantic search is understood as a data retrieval task (e.g., semantic portals)
 - Sometimes it makes partial use of the expressive power of an ontology-based representation

Semantic search: a global classification

Criteria	Approaches
Semantic knowledge representation	Linguistic conceptualization Latent Semantic Analysis Ontology-based information retrieval
Scope	Web search Limited domain repositories Desktop search
Goal	Data retrieval Information retrieval
Query	Keyword query Natural language query Controlled natural language query Structured query based on ontology query languages
Content retrieved	Pieces of ontological knowledge XML documents Text documents Multimedia documents
Content ranking	No ranking Keyword-based ranking Semantic-based ranking

Semantic search: identified limitations

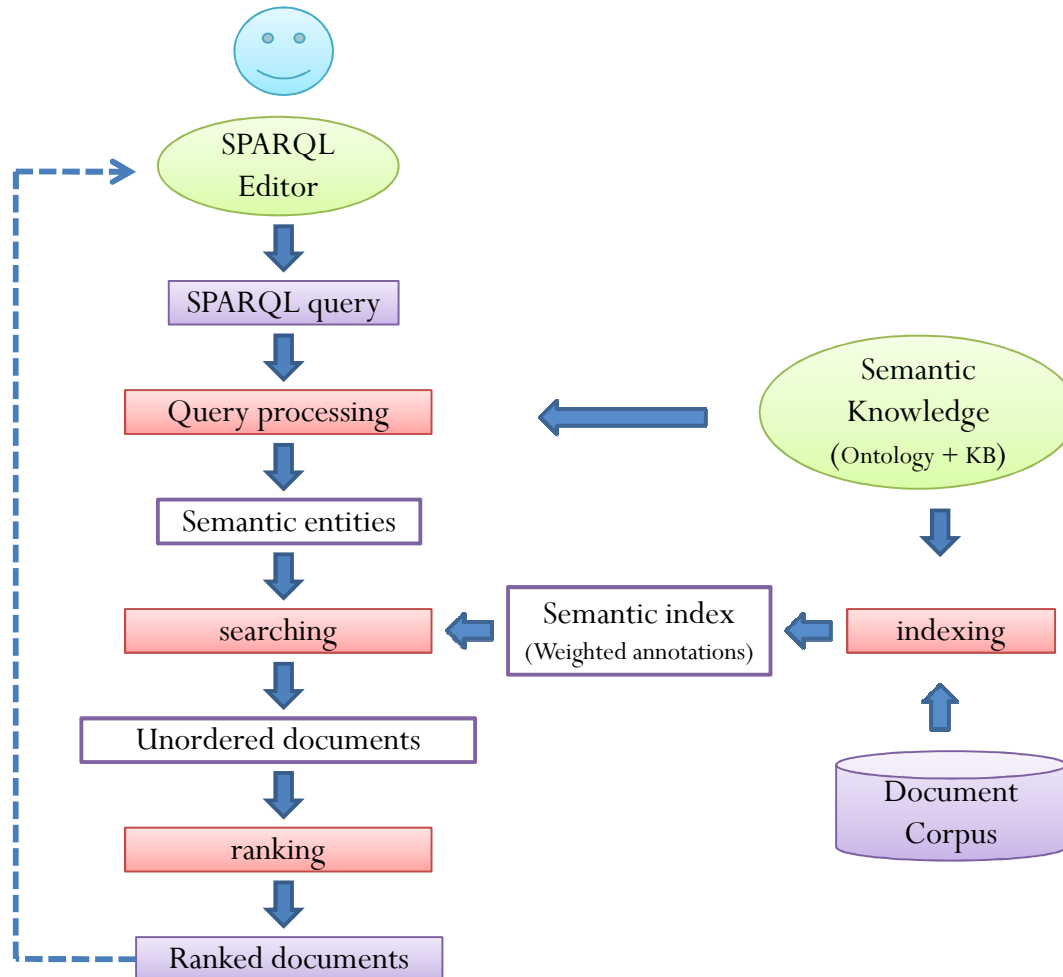
Criteria	Limitations	IR	Semantic
Semantic knowledge representation	Do not exploit the full potential of an ontological language, beyond those that could be reduced to conventional classification schemes	X	(partially)
Scope	Do not scale to large and heterogeneous repositories of documents		X
Goal	Are based on Boolean retrieval models where the information retrieval problem is reduced to a data retrieval task		X
Query	Limited usability		X
Content retrieved	Focused on textual content: unable to manage different formats (multimedia)	(partially)	(partially)
Content ranking	Lack of semantic ranking criteria. The ranking (if provided) relies on keyword-based approaches	X	X
Additional Limitations			
Coverage	Knowledge incompleteness	(partially)	X
Evaluation	Lack of standard evaluation frameworks		X

Part II. The proposal

- Our proposal towards semantic search: an ontology-based IR model
 - Semantic retrieval framework
 - Semantic indexing
 - Query processing
 - Searching and ranking
 - An example
 - Evaluation
 - Results
 - Conclusions
- Semantic retrieval on the Web
 - Limitations of semantic retrieval in the Web environment
 - Semantic retrieval framework extensions
 - Semantic indexing
 - Query processing
 - Searching and ranking
 - Evaluation
 - Results
 - Conclusions

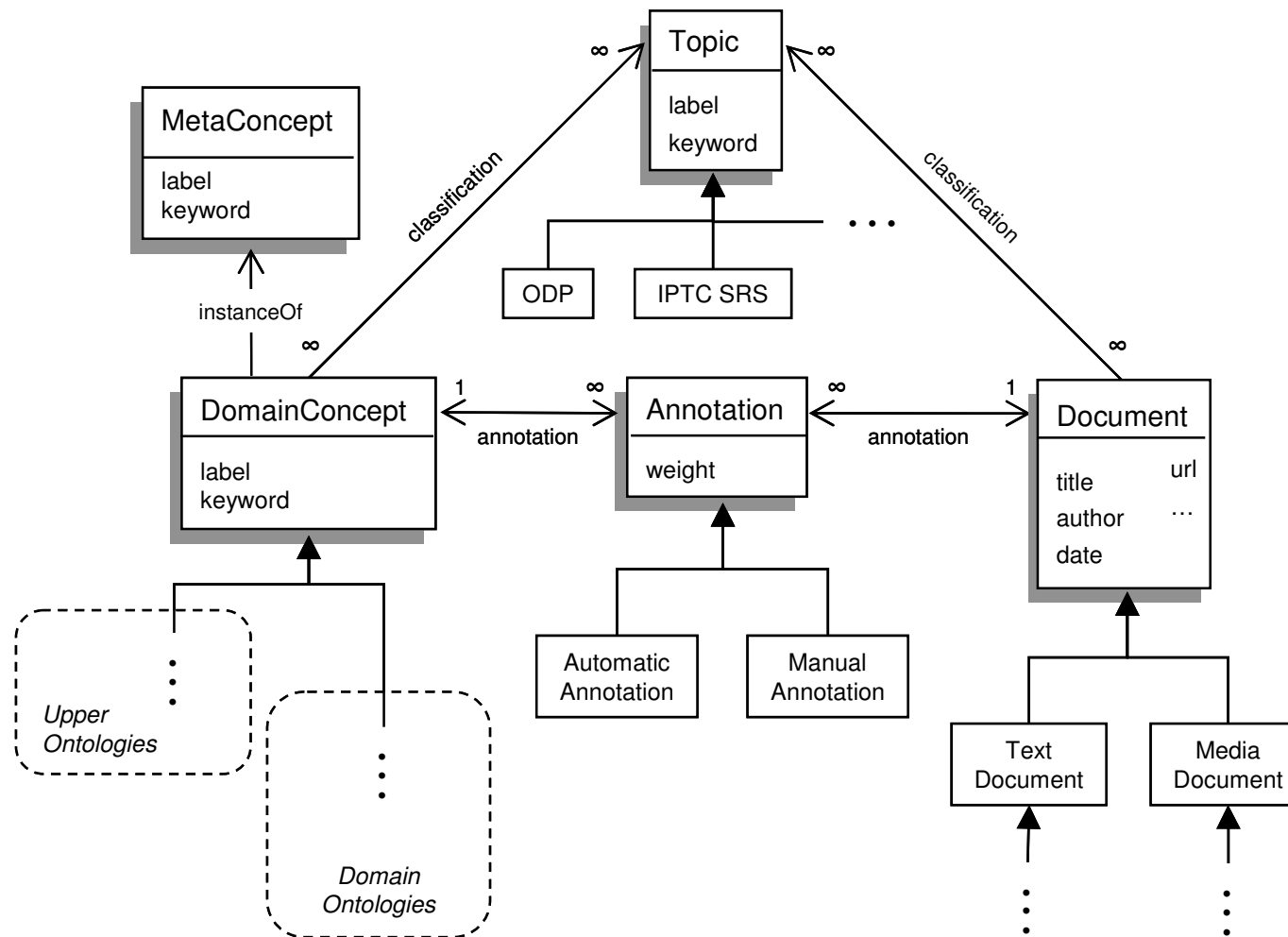
Semantic retrieval framework

- Adaptation of the classic keyword-based IR model
 - Semantic knowledge representation: the “bag of words” is replaced by an ontology and its corresponding KB



Semantic indexing

- Adaptation of the classic inverted index
 - Concepts instead of keywords are associated to documents
 - Annotation weights are computed using an adaptation of the TF-IDF algorithm



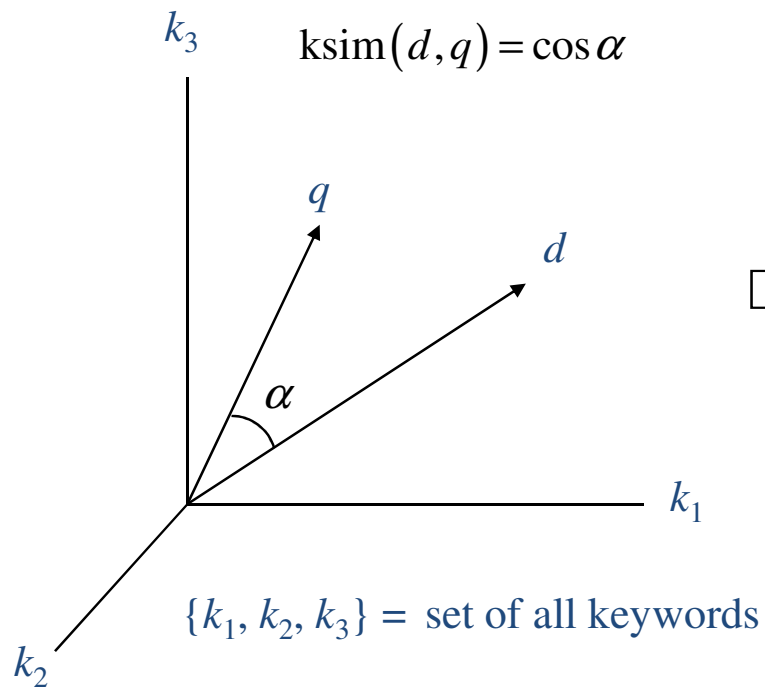
Querying, searching and ranking (I)

- Adapting the vector-space IR model

Keyword-Based IR Model

Query keyword-vector q

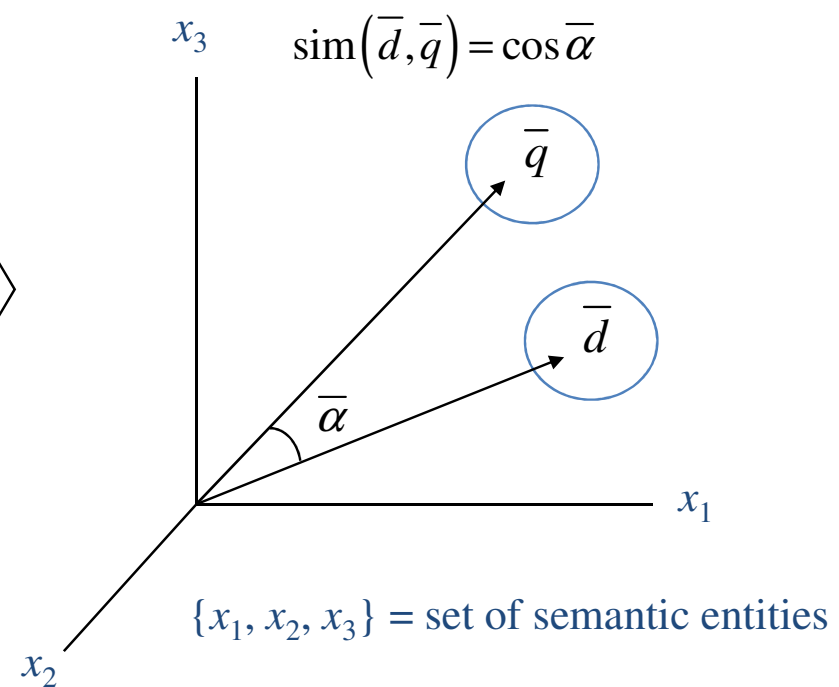
Document keyword-vector d



Semantic IR Model

Result-set concept-vector \bar{q}

Document concept-vector \bar{d}



Querying, searching and ranking (II)

- Building the **query vector** \bar{q}_x
 - Execute the query (e.g. SPARQL) \rightarrow Result set $R \subset O^{|V|}$
 - Variable weighs: for each variable $v \in V$ in the query, $w_v \in [0,1]$
 - For each $x \in O$,
$$\bar{q}_x = \begin{cases} w_v & \text{if } x \text{ instantiates } v \text{ in some tuple in } R \\ 0 & \text{otherwise} \end{cases}$$
- Building the **document vector** \bar{d}_x
 - Map concepts to keywords
 - Weight for an instance $x \in O$ that annotates a document d : TF-IDF

$$\bar{d}_x = \frac{freq_{x,d}}{\max_{y \in O} freq_{y,d}} \cdot \log \frac{N}{n_x}$$

$freq_{x,d}$ = number of occurrences of keywords of x in d

n_x = number of documents annotated by x

N = total number of documents

An example

SPARQL query	Results	
Query: “players from USA playing in basketball teams of Catalonia” PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#> PREFIX kb: http://nets.ii.uam.es/ SELECT ?player ?team WHERE { ?player rdf:type kb:SportsPlayer. ?player kb:plays kb:Basketball. ?player kb:nationality kb:USA. ?player kb:playsIn ?team. ?team kb:locatedIn kb:Catalonia.}	<u>Player</u> (w=1.0) Aaron Jordan Bramlet Derrick Alston Venson Hamilton Jamie Arnold	<u>Team</u> (w=0.5) Caprabo Lleida Caprabo Lleida DKV Joventut DKV Joventut

Query vector: (... ,1, 1, 1, 1, 0.5, 0.5,...)

Found documents: 66 news articles ranked from 0.1 to 0.89. E.g., 1st result

“Johnny Rogers and Berni Tamames went yesterday through the medical revision required at the beginning of each season, which consisted of a thorough exploration and several cardiovascular and stress tests, that their **team** mates had already passed the day before. Both **players** passed without major problems the examinations carried through by the medical **team** of the club, which is now awaiting the arrival of the Northamericans **Bramlett** and **Derrick Alston** to conclude the revisioning.”

Document vector (..., 1.73, ..., 1.65, ...)

Semantic rank value: $\cos(d, q) = 0.88$

Keyword rank value: $\cos(d, q) = 0.06$

Combined rank value: 0.47

Evaluation

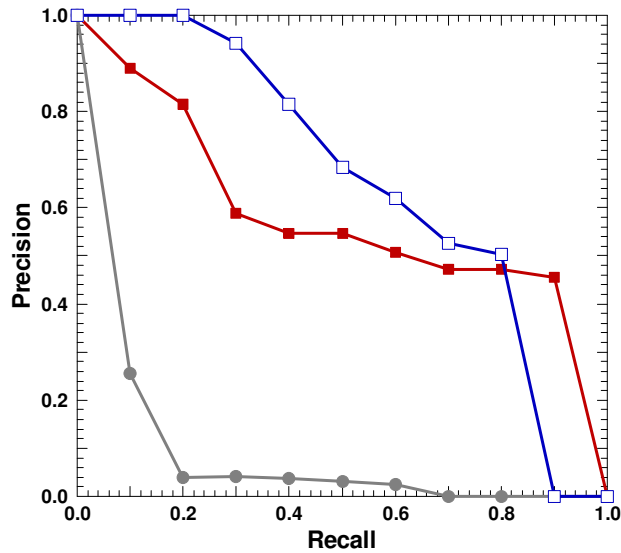
▪ Evaluation benchmark

- Document collection: news articles from the CNN Web Site
 - 145,316 documents (445 MB) from the CNN (NewsArticle → TextDocument)
- Domain ontology and KB: KIM with minor extensions and adjustments
 - 281 domain classes, 138 properties, in several domains
 - 35,689 instances, 465,848 sentences, (71MB in RDF text format)
- Queries
 - A set of twenty queries was prepared manually. $w_v = 1$ for all v in the SPARQL queries
- Judgments:
 - Manual judgement of documents from 0 to 5

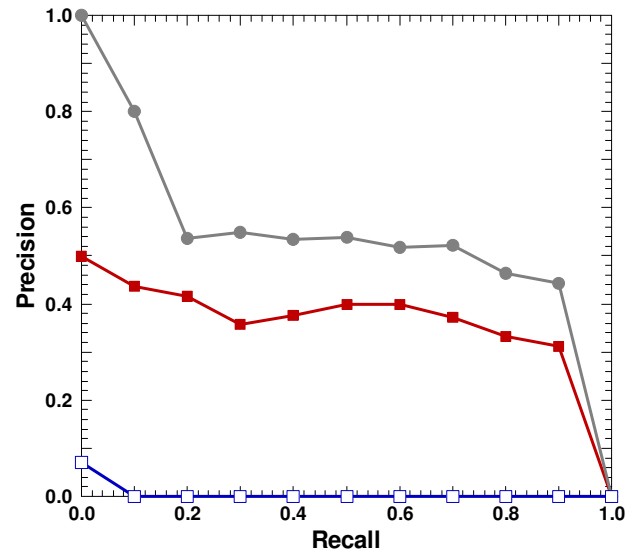
▪ Experimental conditions

- Keyword-based search (Lucene)
- Ontology-only search
- Semantic search

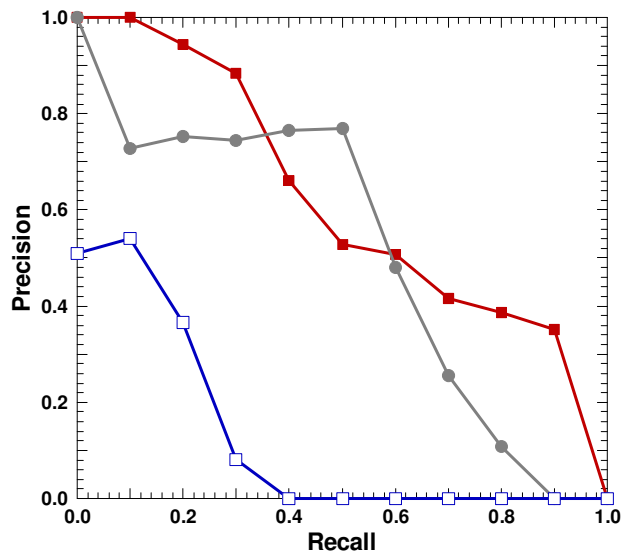
Results



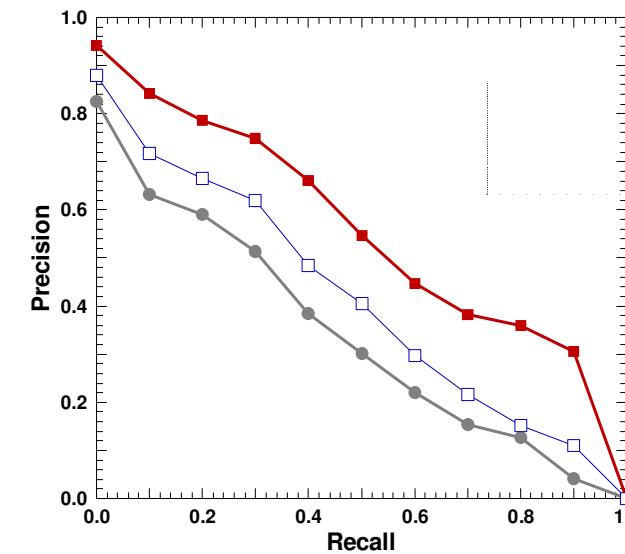
“News about banks that trade on NASDAQ, with fiscal net income greater than two billion dollars.”



“News about insurance companies in USA.”



“News about telecom companies.”



Average

Initial model conclusions

- Better precision by using structured semantic queries (more precise information needs)
 - E.g. a football player **playing in** the Juventus vs. **playing against** the Juventus
- Better recall when querying for instances by class (query expansion)
 - E.g. “News about **companies** quoted on NASDAQ”
- Better recall by using inference
 - E.g. “Watersports in Spain” → ScubaDiving, Windsurf, etc. in Cadiz, Valencia, Alicante, etc.
- Better precision by using query variable weights
 - E.g. new articles about car models released this year, where the **release date** is not necessarily mentioned
- Ambiguity is easier to deal with at the level of concepts
 - Property domain/range, topic-based classification, etc.
- Conditions on concepts and conditions on documents
 - E.g. **film review** published by “Le Monde” within the last 7 days about sci-fi **movie**

Part II. The proposal

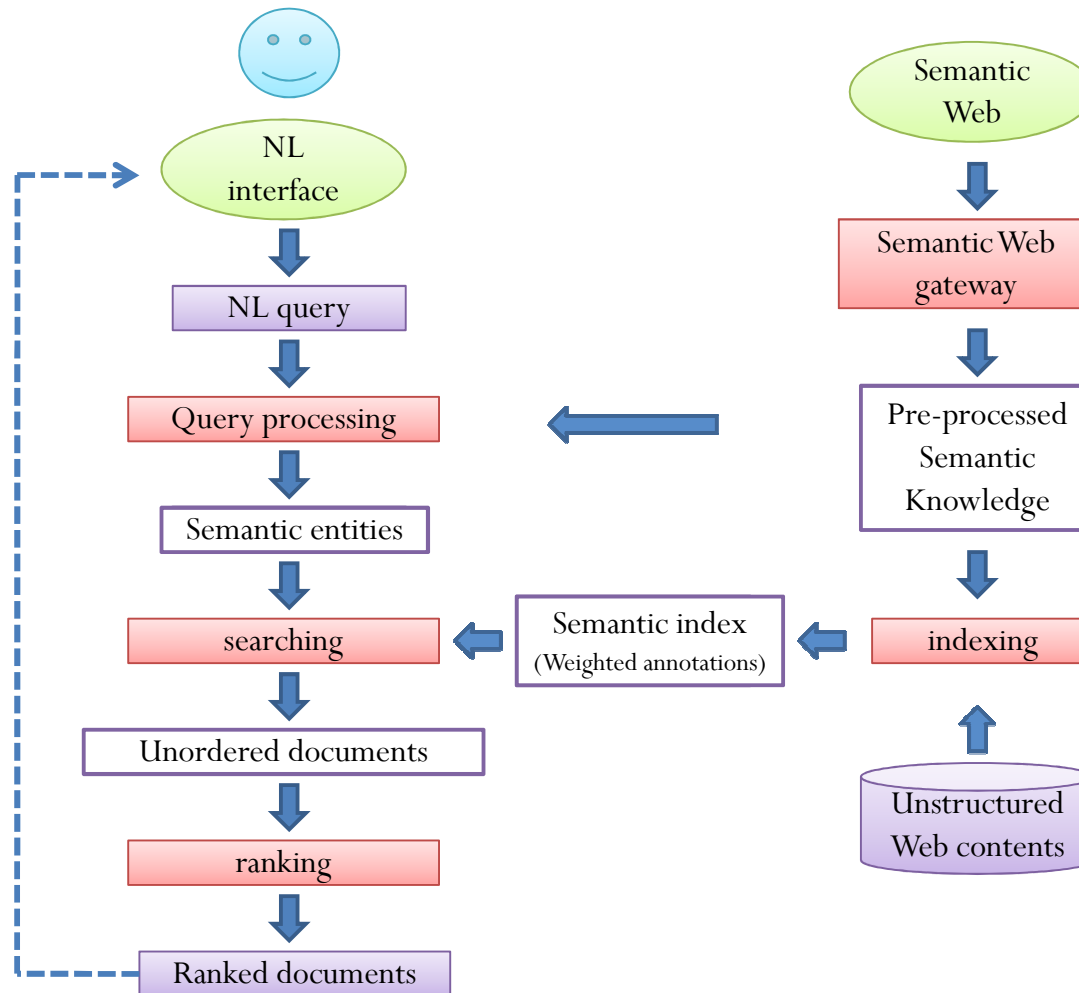
- Our proposal towards semantic search: an ontology-based IR model
 - Semantic retrieval framework
 - Semantic indexing
 - Query processing,
 - Searching and ranking
 - An example
 - Evaluation
 - Results
 - Conclusions
- Semantic retrieval on the Web
 - Limitations of semantic retrieval in the Web
 - Semantic retrieval framework extensions
 - Semantic indexing
 - Query processing
 - Searching and ranking
 - Evaluation
 - Results
 - Conclusions

Limitations of semantic retrieval on the Web

- Applying semantic retrieval on a decentralized, heterogeneous and massive repository of content such as the Web is still an open problem
 - **Heterogeneity:** Web contents span a potentially unlimited number of domains. Impossible to fully cover with a predefined set of ontologies and KBs
 - Proposal: generation of a SW gateway that collects and gives access to semantic metadata available online.
 - **Scalability:** Scaling our model to the Web environment involves exploiting all the semantic metadata available online and to manage huge amounts of information in the form of unstructured content
 - Proposal: creation of scalable and flexible semantic indexing (annotation) methods.
 - **Usability:** Provide users with usable query interface
 - Proposal: support natural language

Semantic retrieval framework extensions

- Queries are expressed in natural language
- A SW gateway is integrated to collect, store and give fast access to the online metadata



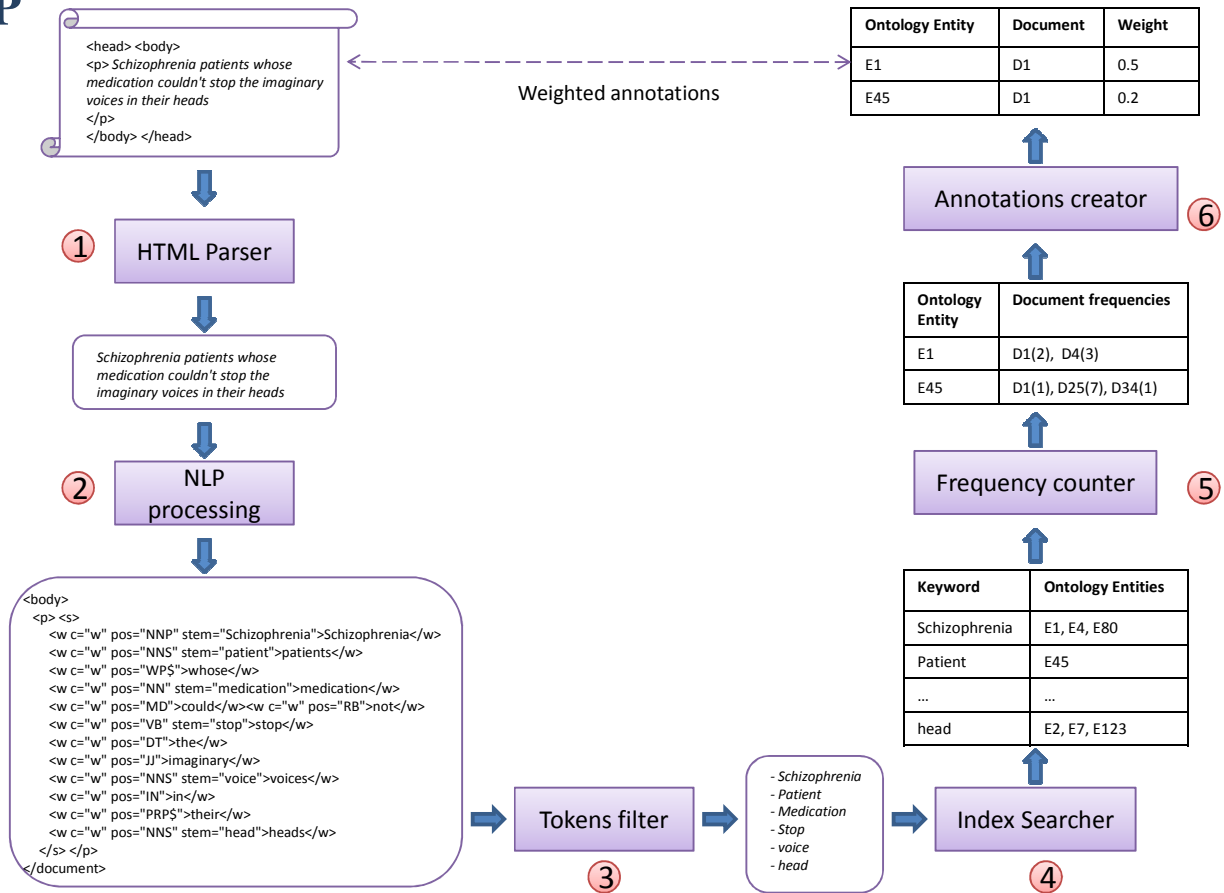
Semantic indexing (I)

- Two different semantic indexing (annotation) methodologies are proposed
 - Annotation based on NLP
 - Annotation based on contextual semantic information
- Common requirements
 - Identify ontology entities (classes, properties, instances or literals) within the documents to generate new annotations
 - Do not populate ontologies, but identify already available semantic knowledge within the documents
 - Support annotation in open domain environments (any document can be associated or linked to any ontology without any predefined restriction) . This brings scalability limitations. To solve them we propose:
 - Generation of ontology indices
 - Generation of document indices
 - Construction of an annotation database which stores non-embedded annotations

Entity ID	Doc Id	Weight
182904817	361452228	0.54
182904817	361452228	0.21

Semantic indexing (II)

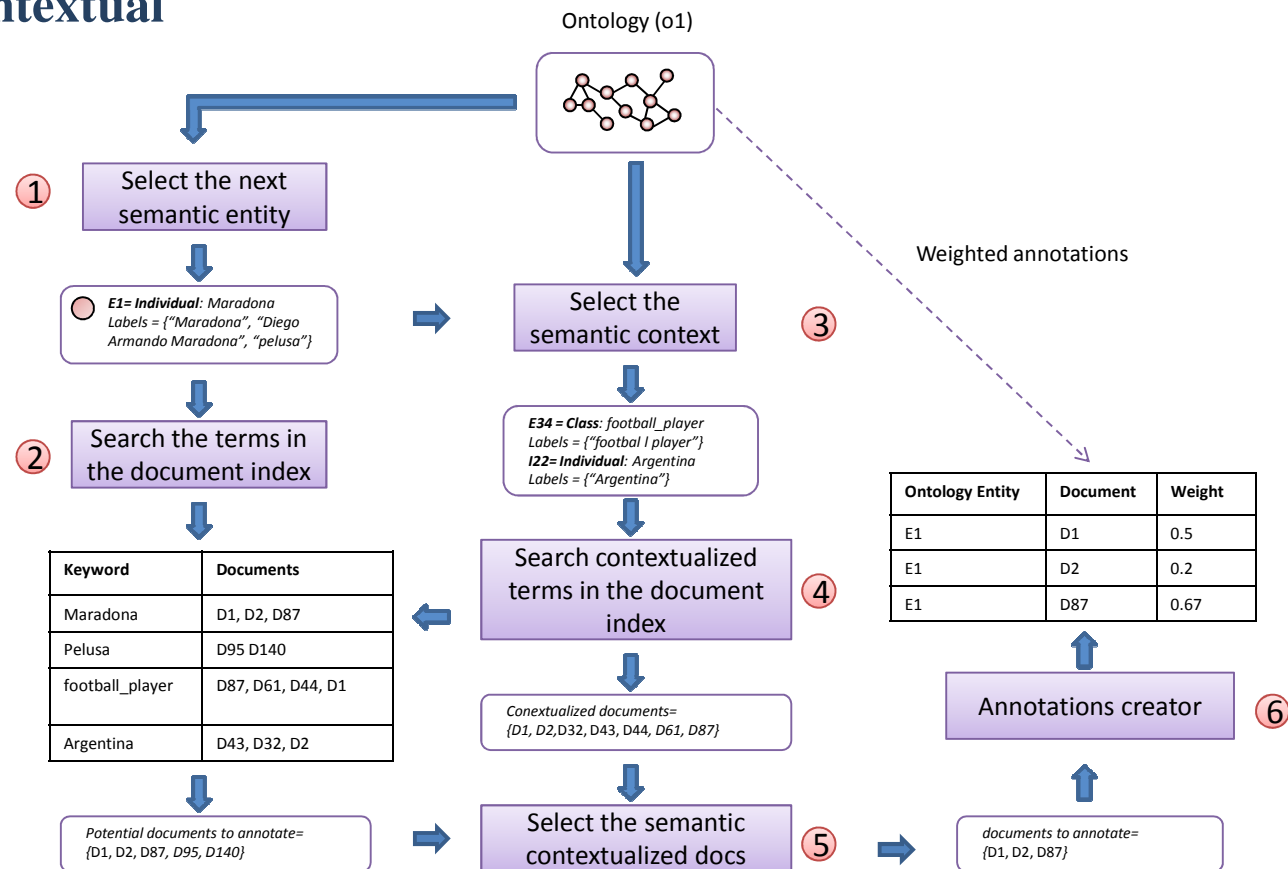
Annotation based on NLP



- **Ambiguities:** exploit the PoS to reduce ambiguities creating groups of words that can potentially express a concept (e.g., *Noun + noun*. “tea cup”)
- **Annotation weights:** use of TF-IDF + PoS to include pronouns when computing frequencies

Semantic indexing (III)

Annotation based on contextual semantic information

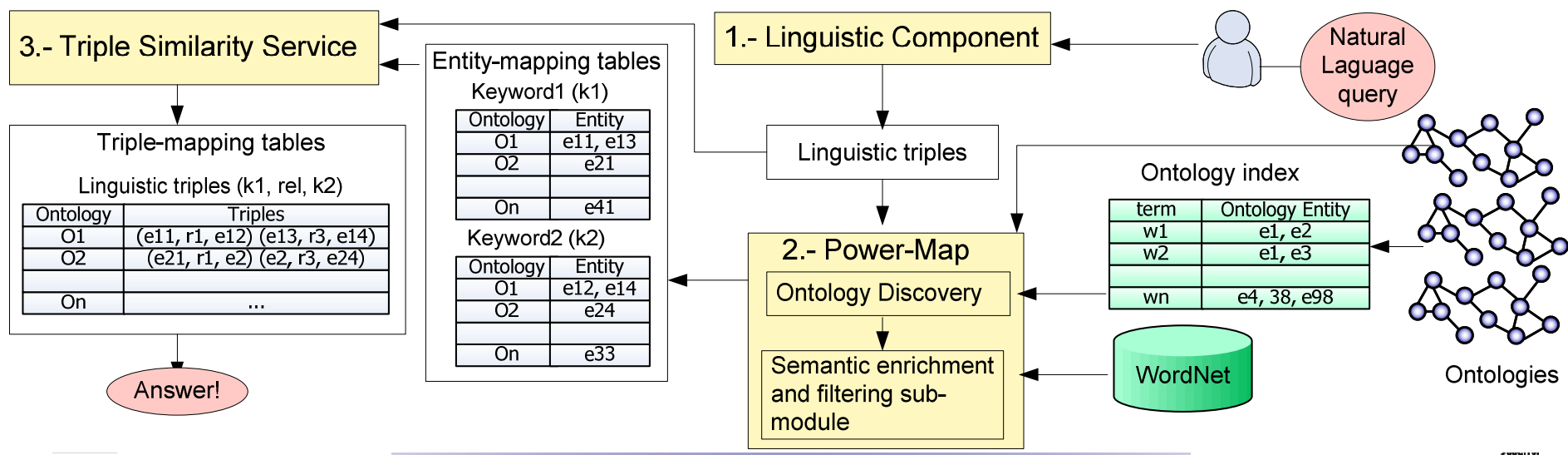


- **Ambiguities:** exploit ontologies as background knowledge (increasing precision but reducing the number of annotations)
- **Annotation weights:** computed from document ranking scores $P \cdot S_d + (1-P) \cdot C_d$

Query processing

- Integration of **PowerAqua** as query processing module

- Input: natural language query
- Output: list of semantic entities retrieved from different ontologies and KBs
- Components
 - Linguistic component: translate the query into its linguistic triple form:
 - “which are the members of the rock group Nirvana?” = *<what-is, members, rock group nirvana>*.
 - PowerMap: maps the terms of each linguistic triple to semantically relevant ontology entities
 - Triple similarity service: selects the ontological triples that best represent the user’s query



Searching and ranking

- Construction of the query vector:
 - The weights are computed considering the set of answers retrieved by PowerAqua
 - The weight of each entity in the query vector is computed as $1/|S|$ where S is the set of semantic entities retrieved for the query condition i
 - E.g., “symptoms and treatments of Parkinson disease” (this query has two conditions)
- Construction of the document vector:
 - Document vectors are computed using semantic entities from different ontologies and KBs

Evaluation

- Evaluation benchmark

- Document collection:
 - TREC WT10G
- Queries and judgments
 - TREC 9 and TREC 2001 test corpora (100 queries with their corresponding judgments)
 - 20 queries selected and adapted to be used by PowerAqua (our QA query processing module)
- Ontologies
 - 40 public ontologies covering a subset of the TREC domains and queries (370 files comprising 400MB of RDF, OWL and DAML)
 - 100 additional repositories (2GB of RDF and OWL) stored and indexed with the SW gateway
- Knowledge Bases
 - Some of the 40 selected ontologies have been semi-automatically populated from Wikipedia

- Experimental conditions

- Keyword-based search (Lucene)
- Semantic-based search
- Best TREC automatic search
- Best TREC manual search

Results (I)

Topic	Semantic	Lucene	TREC automatic	TREC manual
451	0.42	0.29	0.58	0.54
452	0.04	0.03	0.2	0.33
454	0.26	0.26	0.56	0.48
457	0.05	0	0.12	0.22
465	0.13	0	0	0.61
467	0.1	0.12	0.09	0.21
476	0.13	0.28	0.41	0.52
484	0.19	0.12	0.05	0.36
489	0.09	0.11	0.06	0.41
491	0.08	0.08	0	0.7
494	0.41	0.22	0.57	0.57
504	0.13	0.08	0.38	0.64
508	0.15	0.03	0.06	0.1
511	0.07	0.15	0.23	0.15
512	0.25	0.12	0.3	0.28
513	0.08	0.06	0.12	0.11
516	0.07	0.03	0.07	0.74
523	0.29	0	0.23	0.29
524	0.11	0	0.01	0.22
526	0.09	0.06	0.07	0.2
Mean	0.16	0.1	0.2	0.38

MAP: mean average precision

Topic	Semantic retrieval	Lucene	TREC automatic	TREC manual
451	0.7	0.5	0.9	0.8
452	0.2	0.2	0.3	0.9
454	0.8	0.8	0.9	0.8
457	0.1	0	0.1	0.8
465	0.3	0	0	0.9
467	0.4	0.4	0.3	0.8
476	0.5	0.3	0.1	1
484	0.2	0.3	0	0.3
489	0.2	0	0.1	0.4
491	0.2	0.3	0	0.9
494	0.9	0.8	1	1
504	0.2	0.2	0.5	1
508	0.5	0.1	0.3	0.3
511	0.4	0.5	0.7	0.2
512	0.4	0.2	0.3	0.3
513	0.1	0.4	0	0.4
516	0.1	0	0	0.9
523	0.9	0	0.4	0.9
524	0.2	0	0	0.4
526	0.1	0	0	0.5
Mean	0.37	0.25	0.3	0.68

P@10: precision at 10

- Figures in bold correspond to best result for each topic, excluding the best TREC manual approach (because of the way it constructs the query)
- Annotation based on contextual semantic information is used for this experiment

Results (II)

- **By P@10, the semantic retrieval outperforms the other two approaches**
 - It provides maximal quality for 55% of the queries and it is only outperformed by both Lucene and TREC in one query (511)
 - Semantic retrieval provides better results than Lucene for 60% of the queries and equal for another 20%
 - Compared to the best TREC automatic engine, our approach improves 65% of the queries and produces comparable results in 5%
- **By MAP, there is no clear winner**
 - The average performance of TREC automatic is greater than semantic retrieval.
 - Semantic retrieval outperforms TREC automatic in 50% of the queries and Lucene in 75%
 - Bias in the MAP measure
 - More than half of the documents retrieved by the semantic retrieval approach have not been rated in the TREC judgments
 - The annotation technique used for the semantic retrieval approach is very conservative (missing potential correct annotations)

Results (III)

- **For some queries for which the keyword search (Lucene) approach finds no relevant documents, the semantic search does**
 - queries 457 (*Chevrolet trucks*), 523 (*facts about the five main clouds*) and 524 (*how to erase scar?*)
- In the queries in which the semantic retrieval did not outperform the keyword baseline, the semantic information obtained by the query processing module was scarce.
 - Still, overall, **the keyword baseline only rarely provides significantly better results than semantic search**
- TREC Web search evaluation topics are conceived for keyword-based search engines.
 - With complex structured queries (involving relationships), **the performance of semantic retrieval would improve significantly compared to the keyword-based**
 - The full capabilities of the semantic retrieval model for formal semantic queries were not exploited in this set of experiments

Results (IV)

- Studying the impact of retrieved non-evaluated documents
 - **66% of the results returned by semantic retrieval were not judged**
 - P@10 not affected. Results in the first positions have a higher probability of being evaluated
 - MAP: evaluating the impact
 - Informal evaluation of the first 10 unevaluated results returned for every query
 - 89% of these results occur in the first 100 positions for their respective query
 - **A significant portion, 31.5%, of the documents we judged turned out to be relevant**
 - Even though this can not be generalized to all the unevaluated results returned by the semantic retrieval approach (the probability of being relevant drops around the first 100 results and then varies very little) we believe that **the lack of evaluations for all the results returned by the semantic retrieval impairs its MAP value**

Extended model conclusions

- Construction of a complete semantic retrieval approach
 - Input: Natural language queries
 - Output
 - Specific answers in the form of ontology entities
 - Semantically ranked documents
 - Addressing challenges of the Web environment
 - Heterogeneity
 - The system can potentially cover a large amount of domains reusing the ontologies and KBs available online
 - Semantic coverage enhancement would directly result in retrieval performance improvement
 - Scalability
 - The proposed semantic indexing (annotation) methods are able to manage large amounts of unstructured content and semantic metadata without any predefined restriction
 - Need to study in more detail the trade-offs between the quantity and quality of annotations
 - Usability
 - Use of PowerAqua as query processing module. Queries are expressed in NL
 - Knowledge incompleteness
 - If the query processing module does not find any answer, the ranking module ensures that the system degrades gracefully to behave as a traditional keyword-based retrieval approach

Part III. Extensions

- **Semantic knowledge gateway (WebCORE)**
 - Collects, stores and provides access to the semantic content
 - Ontology Indexing module
 - Multi-ontology accessing module
 - Ontology evaluation and selection module
 - Content-based ontology evaluation techniques
 - Collaborative ontology evaluation techniques

- **Coping with knowledge incompleteness**
 - Recall and precision of keyword-based search shall be retained when ontology information is not available or incomplete
 - Making use of rank fusion strategies to combine the results coming from our ontology-based retrieval model and the results returned by traditional keyword-based techniques
 - Proposing a novel score normalization approach based on the behavioral patterns of the search engines (drawn from long-term observations)

Contributions

- Study and **comparison of the different views and approximations to the notion of semantic search** from the IR and semantic technologies fields, identifying fundamental limitations in the state of the art
- Definition, development and formal evaluation of a **novel semantic retrieval model** with deep levels of conceptualization to improve semantic retrieval in large repositories of unstructured information
- **Steps towards semantic retrieval in the Web** environment
- Creation of semantic retrieval **evaluation benchmarks**

Discussion and future work

- Semantic resources
 - Take in **larger amounts of online available semantic metadata** (Watson)
 - Further study on the **trade-off between the quality and quantity of annotations**
- Extensions of the model
 - **Personalization**
 - **Contextualization**
 - **Recommendation**

Thank you!