Using TREC for cross-comparison between classic IR and ontology-based search models at a Web scale

Miriam Fernandez[1], Vanessa Lopez[2], Marta Sabou[2], Victoria Uren[2], David Vallet[1], Enrico Motta[2], Pablo Castells[1]

[1]Escuela Politecnica Superior
Universidad Autonoma de Madrid
C/ Francisco Tomas y Valiente 11,
28048 Madrid, Spain
{Miriam.fernandez, David.vallet, pablo.castells}@uam.es

[2]Knowledge Media institute
The Open University
Walton Hall, Milton Keynes, MK7 6AA,
United Kingdom
{v.lopez, e.motta, r.m.sabou, v.s.uren}@open.ac.uk

ABSTRACT
The construction of standard datasets and benchmarks to evaluate ontology-based search approaches and to compare then against baseline IR models is a major open problem in the semantic technologies community. In this paper we propose a novel evaluation benchmark for ontology-based IR models based on an adaptation of the well-known Cranfield paradigm (Cleverdon, 1967) traditionally used by the IR community. The proposed benchmark comprises: 1) a text document collection, 2) a set of queries and their corresponding document relevance judgments and 3) a set of ontologies and Knowledge Bases covering the query topics. The document collection and the set of queries and judgments are taken from one of the most widely used datasets in the IR community, the TREC Web track. As a use case example we apply the proposed benchmark to compare a real ontology-based search model (Fernandez, et al., 2008) against the best IR systems of TREC 9 and TREC 2001 competitions. A deep analysis of the strengths and weaknesses of this benchmark and a discussion of how it can be used to evaluate other ontology-based search systems is also included at the end of the paper.

Categories and Subject Descriptors
H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval – information filtering, retrieval models, selection process.

General Terms
Measurement, Performance, Experimentation, Standardization.

Keywords
Semantic search, Information Retrieval, evaluation benchmarks.

1. INTRODUCTION
With the continued information explosion, including the emergence of the internet and digital library initiatives, search engines performance has become increasingly critical. In the current commercial competition, designers, developers, vendors and sales representatives of new information products need to carefully study whether and how do their products offer competitive advantages.

This need for search engine performance evaluation has been extensively addressed in the Information Retrieval (IR) research community. As a result several standard evaluation methodologies, metrics and test collections have been developed. The TIPSTER/TREC collections are usually considered to be the reference tests datasets in IR nowadays. The original TIPSTER test design was based on the Cranfield model (Cleverdon, 1967), involving a test collection of documents, user requests (called topics) and relevance judgments. Nowadays different test datasets are built as part of the annual workshops focused on a list of different IR research areas, or tracks, among which we may highlight, for its relevance to our work, the TREC Web track.

In contrast to the IR community, the area of semantic technologies is still a long way from defining standard evaluation benchmarks that comprise all the required information to judge the quality of ontology-based IR approaches. The introduction of ontologies to progress beyond the capabilities of current keyword-based search technologies changes the traditional IR vision in which the user expresses his requirements as a set of keywords and retrieves as answer a ranked set of documents. The most common way in which semantic search has been understood and addressed from the area of semantic-oriented technologies consists of the development of search engines that execute the user query on a KB, and return tuples of ontology values which satisfy the information request (Maedche, Staab, Stojanovic, Studer, & Sure, 2003) (Lopez, Motta, & Uren, 2006). Under such perspective, the document search space is replaced by a semantic search space composed of a set of ontologies and Knowledge Bases (KBs). There are nonetheless works in this context which do explicitly consider keeping, along with the domain ontologies and KBs, the original documents in the retrieval model, where the relation between ontologies and documents is established by annotation relations (Kiryakov, Popov, Terziev, Manov, & Ognyanoff, 2004) (Popov, Kiryakov, Ognyanoff, Manov, & Kirilov, 2004) (Guha, McCool, & Miller, 2003) (Castells, Fernández, & Vallet, 2007).

This difference in the search space used by ontology-based search systems introduces a big gap in the evaluation methodologies used by the two different research communities. While the evaluation methods used by the IR community are systematic, easily reproducible, and scalable, the evaluation methods used by the semantic technologies community rely on user-centered studies (Sure & Iosif, 2002) (McCool, Cowell, & Thurman, 2005) (Todorov & Schandl, 2008) and therefore they tend to be high-cost, non-scalable and difficult to reproduce. This use of user-centered evaluation methods also involves three main limitations:

- The inability to reproduce the experiments and therefore to compare ontology-based search systems against each other.
- The inability to compare ontology-based search systems against traditional IR models using systematic approaches.
- The inability to evaluate ontology-based search systems on a large scale.

This work aims to take a step forward and develop a new reusable evaluation benchmark for cross-comparison between classic IR and ontology-based models on a significant scale. To test the applicability of this benchmark, it is used here to evaluate and compare a specific ontology-based search model (Fernandez, et al., 2008) against available baseline IR systems.

The rest of the paper is structured as follows. Section 2 contains a brief state of the art on evaluation datasets and metrics used by the IR and the semantic technologies communities. Section 3 contains the description of our proposal towards ontology-based evaluation benchmarks. Section 4 presents an example of the application of this proposal to evaluate a specific ontology-based search model (Fernandez, et al., 2008) and analyzes its main strengths and weaknesses. A discussion on how this benchmark can be applied to evaluate other ontology-based search systems is described in Section 5. Finally, conclusions and future work are shown in Section 6.

2. RELATED WORK

As mentioned earlier, while the evaluation of models from the IR community is generally based on systematic approaches, the evaluation of search models from the semantic technologies community is typically user-centered. In this section we briefly describe the evaluation methodologies used by both communities and we analyze which are the main elements needed to develop a common evaluation benchmark.

2.1 Traditional IR evaluation

As we mentioned before, the evaluation of keyword-based retrieval systems is generally based on the Cranfield paradigm (Cleverdon, 1967). In this paradigm, researchers perform experiments on test collections to compare the relative effectiveness of different retrieval approaches using several **evaluation metrics**. The **test reference collection** generally consists of a collection of *documents*, a set of *sample queries*, and a set of relevant documents, *judgments*, manually identified for each query.

2.2 Evaluation metrics

The most common retrieval performance metrics are precision and recall. Consider an example query *q* and its set of relevant documents *R*. Let *A* be the set of documents returned for *q* by a given retrieval strategy under evaluation, and let *Ra* be the documents in the intersection of *R* and *A*, i.e. the relevant documents in the answer set. Recall and precision are defined as:

- **Recall** – is the fraction of the relevant documents which has been retrieved (|Ra|/|R|).
- **Precision** – is the fraction of the retrieved documents which are relevant (|Ra|/|A|).

Note that precision and recall are set-based measures. They evaluate the quality of an unordered set of retrieved documents. To evaluate ranked lists, recall-precision curves are used. For those cases it is common to measure **Precision at 11 standard recall levels**. Each recall-precision point is computed by calculating the precision at the specified recall cutoff value. For the rest of recall values, the precision is interpolated.

As a global estimate of performance across multiple recall levels, it is standard to use **Average Precision (AP)**. This measure is defined as the arithmetic mean of the precision at all the positions in the ranking where a relevant document occurs. To get an average precision of 1.0, a retrieval system must retrieve all relevant documents (i.e., recall = 1.0) and rank them all in the topmost positions, without mix of irrelevant documents (i.e. precision = 1.0 at all positions down to the last relevant document). This measure can be averaged across a set of queries, in which defines the **Mean Average Precision (MAP)**

Beside these basic ones, the list of performance metrics used or proposed in the IR field is considerably large, and is in fact subject to active research. The reader may find in (Baeza Yates & Ribeiro Neto, 1999) an overview on the subject.

2.3 Test reference collections

Test collections are a basic resource for comparing IR systems' performance. Typical text-based collections generally consist of a collection of *documents*, a set of *sample queries*, and a set of relevant documents, *judgments*, manually assigned for each query.

Early attempts at building IR test collections exhaustively judged the relevance of every document to every query. However, for large collections and large numbers of queries (needed to achieve stable and statistically significant measures), providing complete relevance judgements is not feasible. A widely used alternative is *pooled assessment*, in which top-ranked documents from many systems are judged, and unjudged documents are treated as if they were not relevant.

Some of the most popular reference collections nowadays are the ones produced in the TREC initiative. In our present work, we have focused on the **TREC Web track collection used in the TREC 9 and TREC 2001** editions of the TREC conference. The *document collection,* known as WT10g (Bailey, Craswell, & Hawking, 2003), is about 10GB in size, and contains 1.69 million Web pages. It aims to be a testbed for realistic and reproducible experiments on Web documents with traditional, distributed and hyperlink-based retrieval algorithms. The TREC *topics* and *judgements* for this text collection are provided with the TREC 9 and TREC 2001 datasets. Queries were collected from MSN search logs and modified by assessors to meet the TREC topics requirements by adding a description and a narrative to each query (see Figure 1).

<num> 494

<title> nirvana

<desc> Find information on members of the rock group Nirvana

<narr> Descriptions of members' behavior at various concerts and their performing style is relevant. Information on who wrote certain songs or a band member's role in producing a song is relevant. Biographical information on members is also relevant.

Figure 1. Example of a TREC topic

Later on the TREC Web track competitions moved away from the non-Web relevance ranking and towards Web specific tasks, such as finding a particular Web page. Because our work is more focused on the evaluation of ranking quality we did not use for our research the latest available Web track collection.

2.4 Ontology-based search evaluation

As far as the authors are aware there is still little work performed in the formalization of evaluation methodologies for ontology-based search models. One of the main works in this direction is (Sure & Iosif, 2002). In this approach three different search systems were evaluated: QuizRDF and Spectacle, as representative of ontology-based search approaches and EnerSEARCHer as a free text search tool. They designed an experiment with 45 test users divided into 6 groups. The users were provided with a list of 30 questions, 10 for each tool, and were asked to provide, for each question, the number, the answer, the name of the user and the time duration to answer the question. With this information the researchers measured two different things: a) How relatively often did users give wrong, right or no answer with each tool? b) What average relative amount of time needed users for wrong, right or no answers to one single question? To perform these experiments, the authors prepared a set of detailed *questions* and *judgements* in advance. For running the search systems a set of *documents*, a set of *ontologies and KBs*, and a set of *annotations* linking the two search spaces were used.

Another relevant work towards the construction of evaluation benchmarks is (Castells, Fernández, & Vallet, 2007). This benchmark was designed to compare their ontology-based search model against a traditional keyword based approach. For this experiment the authors took as *document collection* 145,316 documents (445 MB) from the CNN Web site. As *ontology and KB* they used the KIM domain ontology and KB (Kiryakov, Popov, Terziev, Manov, & Ognyanoff, 2004), publicly available as part of the KIM Platform; as *annotations* they automatically generated $3 \cdot 10^6$ annotations (i.e. over 25 per document on average) based on the concept-keyword mapping available in the KIM KB; and as *queries and judgements* they manually prepared a set of 20 queries and evaluated the retrieved documents. This work can be considered to achieve a step forward towards the development of ontology-based search evaluation benchmarks on a medium scale.

The work in (Rocha, Schwabe, & Aragão, 2004) presents a qualitative evaluation of their own approach. They prepare a set of *queries* and ask experts to evaluate the quality of the retrieved results (in this case the results are not documents but pieces of ontological knowledge). In total, the experts evaluated 70 different results from two different *ontologies* and *KBs*.

Other evaluation approaches, such as the one reported in (McCool, Cowell, & Thurman, 2005) to test the TAP search engine (Guha, McCool, & Miller, 2003) make use of user-centered evaluation methodologies that evaluate the user satisfaction interacting with the system but do not measure the quality of results returned by the search engine.

3. THE EVALUATION BEHCHMARK

As described in the previous sections, in contrast to traditional IR communities, where standard evaluation methodologies and collections, such as those prescribed by the TREC competitions, have been researched and developed for decades, the semantic technologies community has not yet developed the datasets needed to formally judge the quality of ontology-based search approaches, specially at a large (e.g. Web) scale.

A proper benchmark collection in this area which would meet such requirements should comprise four main components:

- a set of *documents*,
- a set of topics or *queries*,
- a set of relevance *judgments* (or lists of relevant documents for each topic),
- a set of semantic resources, *ontologies and KBs*, which provide the need semantic information for ontology-based approaches.

The described set of components comprises the three main elements used by the Cranfield evaluation paradigm (Cleverdon, 1967) plus the new semantic search space introduced by the ontology-based search approaches.

Starting from this position, and with the aim to move towards a Web-scale evaluation benchmark, we have taken the TREC Web track test corpora as a starting point. In addition, a set of ontologies, KBs, and annotations need to be generated in order to complete the evaluation benchmark and meet the requirements of a semantic-based approach.

For the generation of the semantic search space, the following requirements have been considered: a) in order to approach Web-like conditions, all the semantic search information should be available online; b) the selected semantic information should cover, or partially cover, the domains involved in the TREC set of queries; c) these semantic resources should be completed with a bigger set of random ontologies and KBs to approximate a fair scenario. Given the fact that the semantic resources available online are still scarce and incomplete (Sabou, Gracia, Angeletou, D'Anquin, & Motta, 2007), a fourth requirement has been considered: d) if the semantic information available online has to be extended in order to cover the TREC queries, this must be done with information sources which are completely independent from the document collection, and which are also available online.

Following this set of requirements, the main components of our benchmark are described as follows:

The Document Collection comprises 10 GB of Web documents known as the TREC WT10g collection.

The Queries and Judgements: 20 out of the 100 queries, or topics from TREC 9 and TREC 2001 Web track (corresponding to real user query logs), along with the corresponding relevance judgments, were selected. The selection process is explained in detail in Section 3.2.

The Ontologies: as semantic data on the Web are still sparse and incomplete (Sabou, Gracia, Angeletou, D'Anquin, & Motta, 2007), many of the query topics associated with WT10G are not yet covered by them. Indeed, we only found ontologies covering around 20% of the query topics. In the remaining cases, ontology-based technologies cannot be currently compared against traditional search methodologies, if we are to stick to the fairness requirements stated above. We have thus used 40 public ontologies which potentially cover a subset of the TREC domains and queries. These ontologies are grouped in 370 files comprising 400MB of RDF, OWL and DAML. In addition to the 40 selected ontologies, our benchmark also comprises another 100 repositories (2GB of RDF and OWL).

The Knowledge Bases: sparsity is an even more important problem for KBs than for ontologies. Current publicly available ontologies contain significant structural information in the form of classes and relations. However, most of these ontologies are barely populated or not at all. As a result the available KBs are still not enough to perform significant large-scale experiments . To overcome this limitation, some of the 40 selected ontologies have

been semi-automatically populated from an independent information source: Wikipedia (the population approach is discussed in detail in Section 3.1). Wikipedia is a public encyclopedia comprising knowledge about a wide variety of topics. In this way, we aim to show how semantic information publicly available on the Web can be applied to test ontology-based search approaches over independent sources of documents.

For the ontology-based search approaches that need this information, the benchmark also provides a set of annotations or links between the semantic search space (the ontologies and KBs) and the unstructured search space (the documents). Note that the set of annotations is not strictly necessary to evaluate all the ontology-based search approaches, but it is added to the benchmark in order to enhance its applicability. To generate the annotations we have implemented a completely automatic and scalable approach described in section 3.3.

The annotations: $1.2 \cdot 10^8$ non-embedded annotations have been generated and stored in a MySQL database using the automatic method described in section 3.3.

The final evaluation benchmark thus comprises: a) The TREC WT10g collection of documents; b) 20 queries and their corresponding judgments extracted from the TREC 9 and TREC 2001 competitions; c) 40 public ontologies, some of them populated from Wikipedia, covering the domains of the 20 selected queries, plus 2GB of extra publicly available semantic data, and d) around $1.2\ 10^8$ number of annotations.

3.1 Populating ontologies from Wikipedia

Here we present a simple semi-automatic ontology-population mechanism that can be, in principle, further improved with more sophisticated ontology population techniques, which is out of the extent of this research. The algorithm here comprises two main functionalities: 1) populating an ontology class with new individuals; e.g., populating the class Earthquake with individuals such as 2007 Peru earthquake, 2007 Guatemala Earthquake, etc., and 2) extracting ontology relations for a specific ontology individual, e.g., extract relations for, say, the individual Jennifer Aniston, such as the set of films she has acted in, etc. Basically the algorithm consists of 5 steps:

1. The user selects the class he wants to populate or expand with new relations.
2. The system extracts the textual form of the selected class: either from the localName, or from the standard rdfs:label property, or from some other non-standard ontology property (such as "name", "title", "hasName", etc.) declared as providing a textual form.
3. The system looks for the textual form of the concept in Wikipedia.
4. The contents index of a Wikipedia entry (see Figure 2) is used to generate new classes and/or relations. It suggests sections which point to a list (see Figure 3) or a table (see Figure 4) that can be used to populate the ontology. Note that new classes and relations are created with the information found in the lists and tables only if the algorithm was not previously able to find a mapping in the ontology.
5. The classes selected by the user and the expanded set of classes (in step 4) are populated with the Wikipedia lists and/or tables. To generate a new individual from list entries we take as the individual name the list row up to the first punctuation sign, and the rest of the content as part of the individual rdfs:comment property. To generate a new individual from a table we first create a new class with a set of properties corresponding to the table columns. For each row of the table a new individual of this class is created.

E.g., if we take the entry for the concept Earthquake in Wikipedia, after analyzing the sections pointed to by the contents table shown in Figure 2, the system detects that sections 4, 5 and 6 contain potential lists to populate and extend the concept, and therefore asks the user to select the ones he wishes to exploit. In this case we assume the user selects section 6. The system then analyzes section 6 to generate new classes and properties. First it detects a mapping between "MajorEarthquakes" and "Earthquakes", so it does not create a new class but uses the one in the ontology. For this class the system adds three new subclasses "pre-20 century", "20th century" and "21st century".

Figure 2. Example of Wikipedia contents table

For each subclass the system creates the corresponding instances taking into account the Wikipedia list. The list showed in Figure 3 contains the potential instances for the "Pre-20 century" subclass. After analyzing the first entry of the list the system creates the individual Pompeii and adds the rest of the information "(62)" to the its rdfs:comment property.

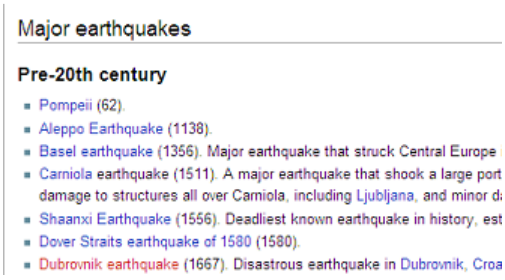

Figure 3. Example of Wikipedia list

With the tables the population process is slightly different, e.g, the table shown in Figure 4 is extracted from the Filmography section of the Jennifer Aniston Wikipedia entrance. For this section the algorithm generates the class "Filmography" with properties: "has year", "has title" and "has role". It also generates the property "hasFilmography" to link the individual "JenniferAniston" with the new "Filmography" individuals created from each row of the table.

Filmography

Year	Title	Role
1990	*Camp Cucamonga*	Ava Schector
1993	*Leprechaun*	Tory Reding
1996	*She's the One*	Renee Fitzpatrick
	Dream for an Insomniac	Allison
1997	*Picture Perfect*	Kate Mosely
	'Til There was You	Debbie

Figure 4 Example of Wikipedia table

This algorithm is supervised. The user identifies the ontology classes to populate, or the ontology instances to extend. He selects from the suggested Wikipedia sections the ones to be used. He can also modify the names of the classes and properties that are automatically generated during the population process.

With this algorithm we have generated around 20.000 triples distributed along the 40 pre-selected ontologies. As said before, this new data added to the KBs has not been extracted from the TREC documents, but from Wikipedia, which maintains the independence assumption for the construction of our benchmark between the semantic search space and the unstructured information search space. It is not our aim to research ontology population methods. Better automatic ontology-population methods could be therefore used to extend the publicly available semantic content with the goal of facilitating ontology-based search approaches such as ours.

3.2 Adapting TREC queries

When selecting the TREC queries to be used in the evaluation benchmark, we had two practical constraints. First, the queries must be able to be formulated in a way that is suitable for ontology-based search systems; for example, queries such as "*discuss the financial aspects of retirement planning*" (topic 514) can not be tackled because they are navigational and not research searches (Guha, McCool, & Miller, 2003). Second, ontologies must be available for the domain of the query. As discussed above, the second point is a serious constraint. Finally, we selected 20 queries.

<num> 494

<title> nirvana

<desc> Find information on members of the rock group Nirvana

<narr> Descriptions of members' behavior at various concerts and their performing style is relevant. Information on who wrote certain songs or a band member's role in producing a song is relevant. Biographical information on members is also relevant.

<adaptation> Show me all members of the rock group nirvana / What are the members of nirvana?

<ontologies> Tapfull, music

Figure 5. Example of a TREC topic.

As shown in Figure 5, TREC queries are described by: a) a title, which is the original user query extracted from users' logs, b) a description and, c) a narrative, which explains in more detail the relevant information that the user is looking for. We added to the queries introduced in the benchmark: d) a detailed request, potentially more suitable for ontology-based search approaches, and e) notes on available ontologies covering that query (see Figure 5). The complete list of the selected TREC topics and their adaptation is available at http://technologies.kmi.open.ac.uk/poweraqua/trec-evaluation.html

3.3 Generating the set of annotations

The overall annotation process is shown in Figure 6, and consists of the following steps to be performed for every semantic entity of each ontology. Note that a standard keyword-based document index is generated prior to the annotation process.

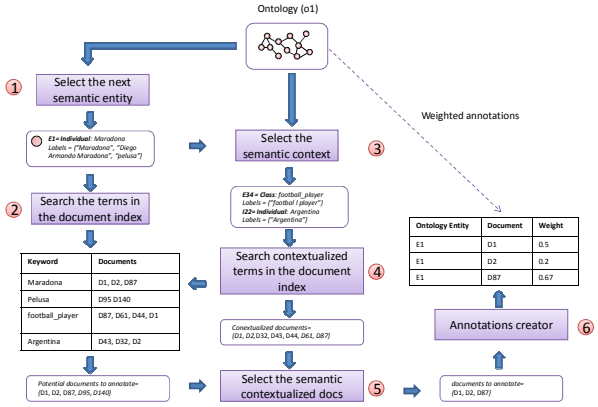

Figure 6. The annotation process

1. *Load the information of a semantic entity*. Extract the textual representation of the selected semantic entity. Each semantic entity has one or more textual representations in the ontology. E.g., the individual entity describing the football player Maradona can be named as "Maradona", "Diego Armando Maradona", "Pelusa", etc. Here we assume that such lexical variants are present in the ontology as multiple values of the local name or rdfs:label property of the entity.
2. *Find the set of potential documents to annotate*. The textual representations of the semantic entity are then searched for in the document index using standard search and ranking processes, in order to find the documents that may be associated with it. These documents simply contain a textual representation of the semantic entity, which does not necessarily imply that they contain its meaning: they are candidates for annotation, to be considered by the following steps.
3. *Extract the semantic context of the entity*. The meaning of a concept is determined by the set of concepts it is linked or related to in the domain ontology. To ensure that a semantic entity annotates the appropriate set of documents, we exploit the ontological relations to extract its context, that is, the set of entities directly linked in the ontology by an explicit relation. E.g., the semantic entity Maradona is related to the concepts Football player, Argentina, etc.
4. *Find the set of contextualized documents*. The textual representations of entities in the set of semantically related concepts, or semantic context, produced in the previous step, are then searched for in the document index to extract the set of contextualized documents.
5. *Select the final list of documents to annotate*. We compute the intersection between the documents having textual representations of the semantic entity (extracted in step 2) and the set of documents having textual representations of the entities on its semantic context (extracted in step 4). Documents in this set

are not just likely to contain the concept but also the contextual meaning of the concept in the ontology.

6. *Create the annotations*. A new entry or annotation is created for every document in the previously obtained set. The annotation will have a weight indicating the degree of relevance of the entity within the document. These weights are computed in the following way: the fusion methodology, described in (Fernandez, 2006), is used on the ranked lists of documents obtained at steps 2 and 4 to produce a ranked list S of documents that are candidates for annotations and a ranked list C of contextualized documents for semantically related entities, respectively. A document d occurring in both, and hence selected for annotation by step 5, will be given a weight $P\,S_d + (1-P)C_d$, where P is a constant used control the influence of the semantic contextualization. We empirically found that a value of $P = 0.6$ seems to work well in practice.

In the annotation mechanism reported here, the semantic entities are analyzed and searched in a standard keyword-base document index. This annotation process provides two important advantages: on the one hand, the semantic information stored in the ontologies and KBs can be used as background knowledge to improve the accuracy of the annotations; on the other hand, this annotation model constitutes a more scalable and widely applicable approach because it can potentially rely on any keyword-based document index, including the ones generated by big search engines.

3.4 Concerns about using the TREC Web track test collection

Several concerns about using the TREC Web track for the evaluation of ontology-based search approaches should be considered:

The judgements: the judgments for each query of TREC 9 and TREC 2001 competitions are obtained using the pooling method described in Section 2. In this methodology, retrieval systems that did not contribute to the pools might retrieve unjudged documents that are assumed to be non-relevant, which, as described in later studies (Voorhees E. , 2001) leads to their evaluation scores being deflated relatively to the methods that did contribute.

The queries: the queries selected for TREC 9 and TREC 2001 are extracted from real Web search engine logs. This means that, the queries are generated in a suitable way for traditional keyword-based search engines and therefore, ontology-based search models are not exploiting their capabilities of addressing more complex types of queries. Consequently, the benchmark might be biased and be giving advantage to keyword-based search approaches.

The query construction: in TREC 9 and TREC 2001 different evaluation categories are considered depending on how the input queries are formulated: a) *short runs,* using just the title or b) *nonshort runs*, automatically or manually constructing the query from the title, the description and the narrative. A better performance is expected from approaches which manually construct the queries (*notshort runs*) than from those that use just the title (*short runs*) because they add a significant amount of additional information to the query. Some semantic search systems could require manually modifying the provided set of queries to fulfill their corresponding input format. However, this does not necessarily mean that additional information is added to the query. In these cases, the comparison of ontology-based search models against traditional IR manual approaches cannot be considered fair.

4. ANALYZING THE EVALUATION BENCHMARK

In order to provide an example of how to use this benchmark and with the more ambitious goal of analyzing its quality, it has been used to compare the results obtained by four different Web scale search approaches. The first three approaches are based on keyword-based retrieval methodologies. The last one is an ontology-based approach.

- **Keyword search**: a conventional keyword-based retrieval approach, using the Jakarta Lucene library.
- **Best TREC automatic search**: the approach used by the best TREC search engine that uses as query just the title section.
- **Best TREC manual search**: the approach used by the best TREC search engine, which manually generates the queries using information from the title, the description and the narrative.
- **Semantic search**: The semantic search system reported in (Fernandez, et al., 2008).

The best TREC search results (title-only and manual) correspond to the best search engines of the TREC 9 and TREC 2001 Web track competitions.

4.1 Results

Table 1 and Table 2 contain the results of our performed evaluation using the 20 TREC topics and the two standard IR evaluation metrics used in the TREC Web track competitions: mean average precision (MAP) and precision at 10 (P@10). The first metric shows the overall performance of a system in terms of precision, recall and ranking. The second one shows how a system works in terms of precision for the top-10 results, which are the ones most likely to be seen by the user.

Numbers in bold correspond to maximal results for the current topic under the current metric, excluding the Best TREC manual approach, which outperforms the others significantly by both metrics likely because of the way the query is constructed: introducing information from the title, the description and the narrative. The other three methodologies construct the query either by using just the title, in the case of the best TREC automatic approach, or by slightly modifying the title to fulfill its corresponding input format in the case the ontology-based search engine. For this reason, **we will exclude Best TREC manual for the rest of our analysis**.

Table 2. Quality of results by MAP

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

Table 2. Quality of results by P@10

Topic	Ontology-based	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

As we can see in Table 2, **by P@10, the ontology-based search outperforms the other two approaches**, providing maximal quality for 55% of the queries. It is only outperformed in one query (511) by both Lucene and TREC automatic. Ontology-based search provides better results than Lucene for 60% of the queries and equal results for another 20%. Compared to the best TREC automatic engine, the semantic approach excels on 65% of the queries and produces comparable results on 5%. Indeed, the highest average value for this metric is obtained by ontology-based search.

The results by MAP are interesting. For those, there is no clear winner. While the average rating for Best TREC automatic is greater than that for ontology-based, ontology-based search outperforms TREC automatic in 50% of the queries and Lucene in 75%.

We hypothesize that the quality of the results retrieved by ontology-based search and its measurement under MAP may be adversely affected by the following factors:

- More than half of the documents retrieved by the ontology-based search approach have not been evaluated in the TREC collection. Therefore, our metrics marked them as irrelevant, when, in fact, some of them are relevant. In Section 4.2 we study the impact of this effect and we manually evaluate some results to analyze how the ontology-based search approach would perform if all documents had been evaluated. The aforementioned section also explains why this factor affects the MAP measurements much more than the P@10 measurements.
- The annotation process used for the semantic retrieval approach is restrictive (see Section 3.3). In order to increase the accuracy of annotations, an annotation is generated when a document contains not just a concept but also its semantic context. If the concept appears in the document with a semantic context not reflected in its ontology, the annotation is not generated. Thus, the process discards possible correct annotations. The trade-offs between the quality and quantity of annotations is another interesting effect whose impact should be analyzed in detail in future experiments.

Another three relevant conclusions can be extracted from this evaluation:

- **For some queries for which the keyword search (Lucene) approach finds no relevant documents, the semantic search does**. This is the case of queries 457 (Chevrolet trucks), 523 (facts about the five main clouds) and 524 (how to erase scar?).
- The queries in which the ontology-based search did not outperform the keyword baseline seem to be those where the semantic information covering the query was scarce. One such query is 467 (Show me all information about dachshund dog breeders). **However, the keyword baseline only rarely provides significantly better results than the ontology-based search**. The effect of the semantic information coverage should be studied in more detail in future work.
- As pointed out before, the effect of complex queries (in terms of relationships) has not been evaluated because TREC Web search evaluation topics are written for keyword-based search engines and do not consider this type of query expressivity. Future work should explore other IR standard evaluation benchmarks such as those used in the QA track, to evaluate the effect of complex queries in the performance of the different search engines. We hypothesize that, under these conditions, **the performance of the ontology-based search approaches would improve significantly relative to that of the others**.

4.2 Impact of retrieved unjudged documents

Given a TREC topic and a document, one of the following three possibilities exists: a) the document is judged as a relevant result; b) the document is judged as an irrelevant result; or c) the document has not been judged in the TREC collection. If semantic search retrieves it, our metrics treat it as irrelevant.

As Table 4 shows, **only 44% of the results returned by the ontology-based search approach had been previously evaluated** in the TREC collection. The unjudged documents, 66%,

are therefore considered irrelevant. However, some of these results may be relevant, and therefore the performance of the ontology-based search approach might be better than reported.

Table 4. Documents retrieved and evaluated by the ontology-based search approach

Topic	Evaluated	Topic	Evaluated
451	44.6%	494	57.3%
452	31.3%	504	32.8%
454	49.4%	508	62.8%
457	54.6%	511	61.3%
465	38.5%	512	39.8%
467	38.0%	513	54.5%
476	50.6%	516	47.5%
484	13.4%	523	20.3%
489	51.6%	524	47.6%
491	47.2%	526	44.6%
Mean			44.4%

Figure 6 shows the probability of a result returned by the ontology-based search approach to be evaluated as function of its position. Results in the first positions have a high probability. In other words, the first results returned by the ontology-based search approach are likely to have also been returned by at least one of the TREC search engines. This explains why unevaluated results are a significant issue for MAP but not for P@10.

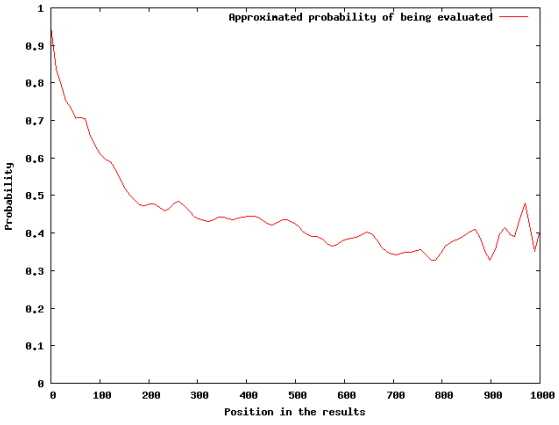

Figure 6. Probability of a document being evaluated by position

We now focus on how the lack of evaluations for documents retrieved by semantic search affects the results for the MAP metrics. A legitimate question is whether the unevaluated results are actually relevant. Indeed, a result is unevaluated if it was not returned by any of the search engines in TREC, which one may expect to imply that it has a low probability of being relevant.

To provide a partial answer to this question we perform an informal evaluation of the first 10 unevaluated results returned for every query, a total number of 200 documents. 89% of these results occur in the first 100 positions for their respective query. We picked the first 10 because these are the most likely to be seen by the user and also because, occurring first on the ranking, they have a larger impact on the MAP measurements.

The results of our evaluation are shown in Table 5. For each query, we show the position in which the 10 documents we evaluated occurred. The positions with a result judged as relevant are shown in bold. We also show the percentage of these results that were judged as relevant.

Table 5. Evaluation of top-10 retrieved unjudge documents

Topic	Positions of top 10 unevaluated results	Rel
451	25,26,27,28,32,34,35,36,37,38	0%
452	2, 4, 5, 6, 7, 9, 10, 18, 20, 21	0%
454	**9, 15,** 22, **38, 42, 43, 49, 56, 61, 63**	90%
457	1, 3, 26, 27, 28, 29, 31, 40, 41, 42	0%
465	4, 5, 8, **10, 16, 21,** 25, **26,** 27, **28**	50%
467	5, **6**, **7**, 12, 13, **14**, 16, **17**, **20**, 28	50%
476	**2**, 3, 7, **11**, 12, **15**, **21**, **23**, 24, 25	50%
484	78, 79, 84, 85, 88, 89, 91, 93, 94, 95	0%
489	**11**, 54, 68, 79, **80**, 82, **83**, 97, 105, 106	30%
491	1, 2, 10, 11, 15, 17, 19, 21, 23, 24	0%
494	**86**, **88**, **128**, 130, 138, 139, **140**, 147, 154, 163	40%
504	**2**, **4**, **5**, **6**, 7, **8**, 9, 11, **12**, 14	60%
508	4, **21**, **22**, 23, **29**, **32**, 39, 41, 48, **52**	50%
511	**4**, 27, **32**, **40**, **42**, **47**, 48, **52**, 60, **61**	70%
512	23, **28**, 30, 31, 33, 35, **63**, 65, 66, **75**	30%
513	61, 62, **76**, 108, 129, **132**, **143**, 150, 153, **157**	40%
516	46, 71, 72, 76, 77, 87, 88, **91**, 96, 100	10%
523	14, 21, 22, **27**, **28**, 29, 37, **41**, 43, 45	30%
524	**0**, 13, 14, **18**, 19, **21**, 50, 59, 60, 61	30%
526	1, 11, 32, 72, 79, 98, 100, 101, 107, 108	0%
Avg:		31,5%

A significant portion, 31.5%, of the documents we judged turned out to be relevant. Clearly, this cannot be generalized to all the unevaluated results returned by the ontology-based search approach: as one moves towards the bottom, the probability of a result being relevant decreases, as shown by Figure 7. This figure is based only on the TREC evaluations, treating unevaluated (by TREC) results as irrelevant, so the actual probability is slightly higher. The figure shows that the probability of being relevant drops around the first 100 results and then varies little. Regardless, we believe that the lack of evaluations for all the results returned by the ontology-based search impairs its MAP value.

The queries for which, of the top-10 documents retrieved that are not evaluated by TREC at least 50% were considered relevant, show that, in most cases, the ontology-based search is obtaining new relevant documents when the query involves a class-instance relationship in the ontologies such as specific symptoms and treatments of Parkinson disease, specific movies or TV programs where Jenifer Anniston appears, etc.

Most of the results in Table 5, even those we consider irrelevant, have related semantic information. For example, for topic 451, although documents about Bengal cats were not retrieved, most of the results were about other types of cats. For topic 457, the results centered around specifications of Chevrolet cars instead of Chevrolet trucks. This "potential recommendation" characteristic of ontology-based search approaches could even have a positive impact on the user's satisfaction, but this should be studied more carefully before definitive conclusions can be drawn.

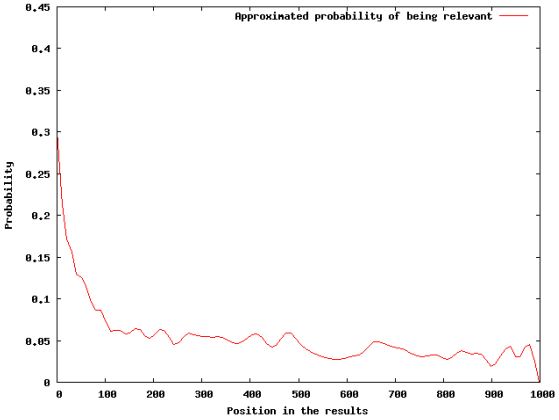

Figure 7. Probability of a document being relevant by position

5. APPLICATIONS OF THE BENCHMARK

In this section we discuss how this benchmark can be applied to evaluate other ontology-based search approaches. As opposed to traditional IR approaches, that use the same type of inputs (queries) and outputs (ranked documents), nowadays there is no standard model of ontology-based search. The different models described in the literature present different inputs, outputs and scope. A brief summary of these differences is shown in Table 6.

Table 6. Classification of ontology-based search systems

Criteria	Approaches
Scope	Web search Limited domain repositories Desktop search
Input (query)	Keyword query Natural language query Controlled natural language query Ontology query languages
Output	Data retrieval Information retrieval
Content ranking	No ranking Keyword-based ranking Semantic-based ranking

Scope: the application of semantic search has been undertaken in different environments such as *the Web* (Finin, Mayfield, Fink, Joshi, & Cost, 2005), *Controlled Repositories* generally restricted to a predefined set of domains (Popov, Kiryakov, Ognyanoff, Manov, & Kirilov, 2004) or even the Desktop (Chirita, Gavriloaie, Ghita, Nejdl, & Paiu, 2005).

Input (query): another relevant aspect that characterizes semantic search models is the way the user expresses his information needs. Four different approaches may be identified in the state of the art, characterized by a gradual increase of their level of formality and usage complexity. In the first level, queries are expressed by means of *keywords* (Guha, McCool, & Miller, 2003). For instance, a request of information about movies where Brad Pitt plays the leading role could be expressed by a set of keywords like "Brad Pitt movies". The next level involves a natural *language representation* of the information need (Lopez, Motta, & Uren, 2006). In this case, the previously mentioned example could be expressed as a full (interrogative) sentence, such as "in what movies does Brad Pitt play the leading role?" The next level is represented by *controlled natural language systems* (Bernstein & Kaufmann, 2006) where the query is expressed by adding tags that represent properties, values or objects within the consultation. "s: Actor p: name v: Brad Pitt p: leading-role s: film". Finally the most formal search systems are based on ontology-query languages such SPARQL (Castells, Fernández, & Vallet, 2007), etc. In this approach, the previous example could be expressed as "select ?f where (?a , < name>, 'Brad Pitt') , (?a, <leading-role>, ?f)"

Ouput: ontology-based search approaches can be characterized by whether they aim at data retrieval or information retrieval (IR). While the majority of IR approaches always return documents as response to user requests, and therefore should be classified as information retrieval models, a large amount of ontology-based approaches return ontology instances rather than documents, and therefore may be classified as data retrieval models. E.g., as a response to the query "films where Brad Pitt plays the leading role" a data retrieval system will retrieve a list of movie instances while an IR system will retrieve a list of documents containing information about such movies. Semantic Portals (Maedche, Staab, Stojanovic, Studer, & Sure, 2003) and QA systems (Lopez, Motta, & Uren, 2006), typically provide simple search functionalities that may be better characterized as semantic data retrieval rather than information retrieval.

Content ranking: The definition of ranking in ontology-based search models is currently an open research issue. Most approaches do not consider ranking query results in general; other models base their ranking functionality on traditional keyword-based approaches (Guha, McCool, & Miller, 2003) and a few take advantage of semantic information to generate query result rankings. Generally, KB instances rather than documents are ranked (Stojanovic, 2003).

Considering these different views of the ontology-based search paradigm, the question is: how the constructed benchmark (Section 3) can be applied to evaluate other ontology-based search approaches as well?

If the scope of the ontology-based search model is not a heterogeneous environment such as the Web but a predefined set of domains, a potential solution to apply this benchmark will be to select the set of queries related with the domain where the evaluation should be performed. If there is no a significant amount of queries available in the TREC Web track to perform a suitable comparison, the benchmark cannot be applied as is and should be extended or modified.

Regarding the different types of queries, ontology-based search approaches can be divided into two different groups as well as TREC does with the different IR models (the *short* and *non-short* categories). Ontology-based approaches can be divided among those that do not use structured information in the query (keyword and natural language) and those that use structured queries (controlled natural language queries and ontology query languages).

For the last category, the queries currently provided by the benchmark should be adapted.

Regarding the type of output, the approaches that retrieve documents as answers can be evaluated using this benchmark. For the data retrieval approaches, one potential way of performing their evaluation is to consider their answers as a kind of query expansion. The expanded query is then used on the document space using a traditional keyword-based search. An example of such an evaluation using this benchmark has been performed for PowerAqua (Lopez, Motta, & Uren, 2006) and it is presented in (Fernandez, et al., 2008). It can be argued that, the evaluation of data retrieval approaches using this methodology does not evaluate their real contribution. In this sense, the comparison of ontology-based search systems against traditional Question Answering (QA) models should be more adequate.

Regarding the ranking, not all the ontology-based search approaches retrieve ranked answers. For those cases, evaluation measures such as precision and recall can still be used to perform a more informal evaluation.

6. CONCLUSIONS AND FUTURE WORK

As we have discussed in the previous sections there is a general necessity in the semantic search community to construct standard evaluation benchmarks to evaluate and compare ontology-based approaches against each other and against traditional IR models.

Aiming to advance this issue, we have constructed a potentially widely applicable ontology-based evaluation benchmark departing from traditional IR datasets, such as the TREC Web track reference collection. The model has been used to evaluate a specific ontology-based search approach (Fernandez, et al., 2008) against different traditional IR models at a large scale.

Potential limitations of this benchmark are: a) the need of ontology-based search systems to participate in the pooling methodology to obtain a better set of document judgments, b) the use of queries with a low level of expressivity in terms of relations, more oriented to traditional IR models and, c) the scarceness of the publicly available semantic information to cover the meanings involved in the document search space.

However, despite these limitations, this benchmark constitutes a first step in the evaluation of ontology-based search approaches against traditional IR standards on a Web scale. Its potential application to other ontology-based search approaches has also been analyzed. As a conclusion, we can say that, a common understanding of ontology-based search in terms of inputs, outputs and scope should be reached before achieving a real standardization in the evaluation of ontology-based search models.

7. REFERENCES

[1] Baeza Yates, R., & Ribeiro Neto, B. (1999). Modern Information Retrieval. Harlow, UK: Addison-Wesley.

[2] Bailey, P., Craswell, N., & Hawking, D. (2003). Engineering a multi-purpose test collection for web. Information Processing and Managment , 853-871.

[3] Bernstein, A., & Kaufmann, E. (2006). Gino - a guided input natural language ontology editor. 5th International Semantic Web Conference. Athens, GA, USA: Springer Verlag.

[4] Castells, P., Fernández, M., & Vallet, D. (2007). An Adaptation of the Vector-Space Model for Ontology-Based Information Retrieval. IEEE Transactions on Knowledge and Data Engineering 19(2), Special Issue on Knowledge and Data Engineering in the Semantic Web Era , 261-272.

[5] Chirita, P. A., Gavriloaie, R., Ghita, S., Nejdl, W., & Paiu, R. (2005). Activity based metadata for semantic desktop search. 2nd European Semantic Web Conference. Heraklion, Greece.

[6] Cleverdon, C. (1967). The Cranfield tests on index language devices. Aslib Proceedings , 173-192.

[7] Fernandez, M., Vallet, D., Castells, P. Probabilistic Score Normalization for Rank Aggregation. 28th European Conference on Information Retrieval (ECIR 2006). London, UK, April 2006. Springer Verlag Lecture Notes in Computer Science, Vol. 3936, ISBN 3-540-33347-9, 2006, pp. 553-556

[8] Fernandez, M., Lopez, V., Sabou, M., Uren, V., Vallet, D., Motta, E., et al. (2008). Semantic Search meets the Web. 2nd IEEE International Conference on Semantic Computing (ICSC 2008). Santa Clara, CA, USA.

[9] Finin, T., Mayfield, J., Fink, C., Joshi, A., & Cost, R. S. (2005). Information retrieval and the semantic Web. 38th Annual Hawaii international Conference on System Sciences (Hicss'05), 4.

[10] Guha, R. V., McCool, R., & Miller, E. (2003). Semantic search. 12th International World Wide Web Conference (WWW 2003), (pp. 700-709). Budapest, Hungary.

[11] Kiryakov, A., Popov, B., Terziev, I., Manov, D., & Ognyanoff, D. (2004). Semantic Annotation, Indexing, and Retrieval. Journal of Web Semantics 2, Issue 1 , 49-79.

[12] Lopez, V., Motta, E., & Uren, V. (2006). PowerAqua: Fishing the Semantic Web. European Semantic Web Conference. Montenegro.

[13] Maedche, A., Staab, S., Stojanovic, N., Studer, R., & Sure, Y. (2003). SEmantic portAL: The SEAL Approach. Spinning the Semantic Web. MIT Press , 317-359.

[14] McCool, R., Cowell, A. J., & Thurman, D. A. (2005). End-User Evaluations of Semantic Web Technologies. Workshop on End User Semantic Web Interaction. ISWC 2005. Galway, Ireland.

[15] Popov, B., Kiryakov, A., Ognyanoff, D., Manov, D., & Kirilov, A. (2004). KIM – A Semantic Platform for Information Extraction and Retrieval. Journal of Natural Language Engineering 10, Cambridge University Press , 375-392.

[16] Rocha, C., Schwabe, D., & Aragão, M. P. (2004). A Hybrid Approach for Searching in the Semantic Web. 13th International World Wide Web Conference (WWW 2004), (pp. 374-383). NY.

[17] Stojanovic, N. (2003). On Analysing Query Ambiguity for Query Refinement: The Librarian Agent Approach. 22nd International Conference on Conceptual Modeling. 2813, pp. 490-505. Berlin Heidelberg: Springer Verlag.

[18] Sure, Y., & Iosif, V. (2002). First Results of a Semantic Web Technologies Evaluation. Common Industry Program at the federated event: ODBASE'02 Ontologies, Databases and Applied Semantics. California, Irvine.