Web Semantics: Science, Services and Agents on the World Wide Web 30 (2015) 52–68

Contents lists available at ScienceDirect

Web Semantics: Science, Services and Agents on the World Wide Web

journal homepage: www.elsevier.com/locate/websem

Mímir: An open-source semantic search framework for interactive information seeking and discovery

Valentin Tablan, Kalina Bontcheva *, Ian Roberts, Hamish Cunningham

University of Sheffield, Department of Computer Science, Regent Court, 211 Portobello, S1 4DP, Sheffield, UK

ARTICLE INFO

Article history:
Received 15 July 2013
Received in revised form
7 October 2014
Accepted 13 October 2014
Available online 22 October 2014

Keywords:
Natural language processing
Semantic search
Scalable semantic search framework
Expressive semantic queries
Integrated semantic search

ABSTRACT

Semantic search is gradually establishing itself as the next generation search paradigm, which meets better a wider range of information needs, as compared to traditional full-text search. At the same time, however, expanding search towards document structure and external, formal knowledge sources (e.g. LOD resources) remains challenging, especially with respect to efficiency, usability, and scalability.

This paper introduces Mímir—an open-source framework for integrated semantic search over text, document structure, linguistic annotations, and formal semantic knowledge. Mímir supports complex structural queries, as well as basic keyword search.

Exploratory search and sense-making are supported through information visualisation interfaces, such as co-occurrence matrices and term clouds. There is also an interactive retrieval interface, where users can save, refine, and analyse the results of a semantic search over time. The more well-studied precision-oriented information seeking searches are also well supported.

The generic and extensible nature of the Mímir platform is demonstrated through three different, real-world applications, one of which required indexing and search over tens of millions of documents and fifty to hundred times as many semantic annotations. Scaling up to over 150 million documents was also accomplished, via index federation and cloud-based deployment.

1. Introduction

Traditional full-text search is no longer able to address the more complex information seeking behaviour, which has evolved towards sense-making and exploratory search [1]. In the latter cases, traditional precision-oriented approaches from the field of Information Retrieval (IR) are not sufficient. For exploratory search, in particular, recall is paramount, as well as the ability to carry out interactive retrieval [1].

Semantic search over documents aims to address these new challenges by finding information that is not based just on the presence of words, but also on their meanings [2]. It is often referred to as *hybrid* or *semantic full-text search* [3], in order to distinguish it from semantic web search engines, concept search and other types of semantic search (see Section 6 for details). Such systems support hybrid semantic queries, which combine keywords and formal query syntax (e.g. SPARQL [4]), in order to search jointly against document content and ontologies.

Semantic full-text or hybrid search is a modification of classical IR, where documents are retrieved on the basis of relevance to ontology concepts, as well as words. While the basic IR approach considers word stems as tokens, there has been considerable effort towards using word-senses or lexical concepts (see [5,6]) for indexing and retrieval. In the case of semantic search, what is being indexed is typically a combination of words, formal knowledge typically expressed in an ontology, and semantic annotations mentioning ontological concepts in the text [2].

Natural Language Processing (NLP) is commonly used to derive semantics from unstructured content and to encode it in a structured format, suitable for semantic search. Since some of the most frequently used searches are for persons, locations, organisations, and other named entities [7], some of the most widely used NLP techniques are *named entity recognition* [8,9], *entity linking or disambiguation* [10], and other types of *semantic annotation* [11].

From a retrieval perspective, entity-annotated content enables semantic search queries such as "LOC earthquake" which would return all documents mentioning a location of an earthquake. Semantic annotation, on the other hand, goes one step further by disambiguating which specific real-world location is mentioned in the text (e.g. Cambridge, UK vs. Cambridge, MA, USA). Typically a knowledge base or a Linked Open Data (LOD) resource is used as

* Corresponding author.
E-mail address: k.bontcheva@sheffield.ac.uk (K. Bontcheva).

http://dx.doi.org/10.1016/j.websem.2014.10.002
1570-8268/

a source of unique entity identifiers (URIs) and formal knowledge about them. This enables even more powerful semantic searches, based on knowledge that is external to document collections. For example, a query on flooding in the UK would retrieve a document about floods in Cambridge, even though the latter does not explicitly mention the UK. The knowledge linking Cambridge to the UK would instead come from, e.g. DBpedia [12] or Geonames.[1]

The focus of this paper is on Mímir[2]—an open-source framework for integrated semantic search over text, document structure, linguistic annotations, and formal semantic knowledge.

Typically semantic full-text search approaches enlarge standard IR indexes with semantic terms (e.g. URIs), while still modelling documents as bags of tokens and disregarding their structure. In contrast, the Mímir semantic search framework uses two additional types of data: *linguistic annotations* created by NLP tools (e.g. morphology, part-of-speech, and syntax) and *document structure annotations* (e.g. paragraphs, sections, titles). In order to distinguish this from the bag-of-words-based semantic full-text search approaches, the term *integrated semantic search* is introduced.

The novelty of the Mímir semantic search framework lies in its support for serendipitous information discovery tasks, to complement information seeking searches. Exploratory search and sense-making are supported through a number of visualisations, including co-occurrence matrices and term clouds, as well as an interactive retrieval interface, where users can save, refine, and analyse the results of a semantic search over time. The more well-studied precision-oriented information seeking searches are also supported, including ranking of search results. To the best of our knowledge, the Mímir framework is the first open-source semantic search platform of this kind.

The novel contributions of this paper are:

1. An in-depth description of the Mímir open-source framework, including its architecture (Section 2), the indexing and search over document text, structure, linguistic annotations, and formal semantic knowledge (Sections 2.1 and 2.2 respectively). In particular, direct indexes are created in addition to the widely used inverted indexes, in order to support both information discovery and information seeking searches. Direct indexes power the dynamic calculation of sets of frequently occurring terms within relevant document lists. These term sets underpin user interfaces that support the discovery of new knowledge and relationships by displaying term clouds and co-occurrence matrices as part of interactive retrieval tasks.
2. Presentation of two semantic search interfaces for information seeking tasks from two real-world applications (Section 3.1).
3. Presentation of a semantic search interface for information discovery, including a real-world application in knowledge discovery from immunology literature (Section 3.2). 9.5 million documents are searched interactively, in conjunction with a medical domain ontology.
4. A comprehensive evaluation of the Mímir semantic search framework. First, intrinsic evaluation is carried out, with respect to indexing and search efficiency (Section 5.1). This includes also the evaluation of a complex semantic search query against 9.5 million documents, 3.5 billion tokens, and 743 million linguistic and semantic annotations (Section 5.2). Second, extrinsic evaluation is carried out with users, as part of a semantic search application which combines environmental science literature and Linked Open Data (Section 5.3).
5. Positioning Mímir with respect to the state-of-the-art (Section 6), including a detailed comparison against the Broccoli semantic full-text search system, which is its nearest analogue.

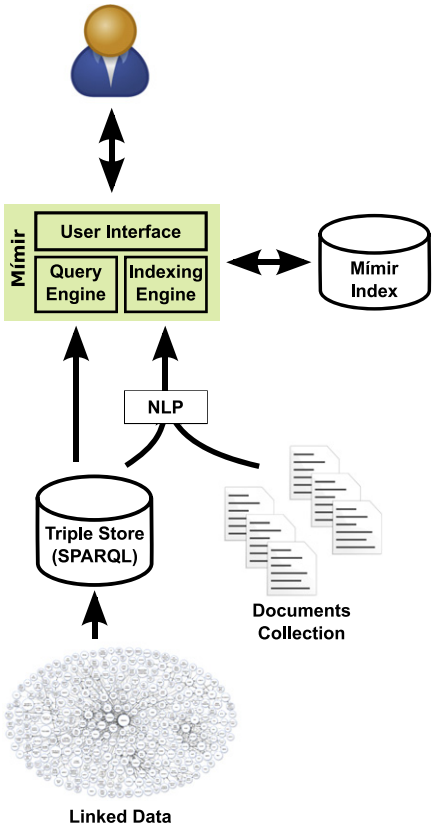

Fig. 1. Mímir life cycle.

2. Mímir: an open-source semantic search framework

Mímir[3] is an integrated semantic search framework, which offers indexing and search over full text, document structure, document metadata, linguistic annotations, and any linked, external semantic knowledge bases. It supports hybrid queries that arbitrarily mix full-text, structural, linguistic and semantic constraints. A key distinguishing feature is the containment operators that allow flexible creation and nesting of full-text, structural, and semantic constraints, as well as Mímir's support for interactive knowledge discovery.

Mímir has been designed as a generic and extensible, open source framework.[4] It can also be used as an on-demand, highly scalable semantic search server, running on the GATECloud [13] platform.

The high-level concept behind Mímir is illustrated in Fig. 1. First a document collection is processed with NLP algorithms, such as those provided by GATE [14]. Typically the semantic annotations also refer to Linked Open Data resources, accessed via a triple store, such as OWLIM [15] or Sesame [16]. The semantically annotated documents are then indexed in Mímir, together with their full-text content, document metadata, and document structure markup. At search time, the triple store is used as a source of implicit knowledge, to help answer the hybrid searches that combine full-text, structural, and semantic constraints. The latter are formulated using a SPARQL query, executed against the triple store.

Mímir's architecture is shown in Fig. 2. It is implemented as a web application that runs server-side and can optionally be distributed across multiple machines. It includes both information

[1] A geographical database available from http://geonames.org.
[2] http://gate.ac.uk/mimir/.
[3] Old Norse "The rememberer, the wise one".
[4] Download from http://gate.ac.uk/mimir/.