Chapter 6
Shared Ontology for Knowledge Management

Atanas Kiryakov, Borislav Popov, Ilian Kitchukov, and Krasimir Angelov

Abstract This chapter focuses on semantic searching at web scale. The solution presented takes advantage of the specific strengths of semantic repositories and the raw power of relational databases, the latter having been developed over decades and capable of handling efficiently large volumes of data with fixed structure (which is the case with the occurrence statistics) and the former allowing for inference and querying on top of formal knowledge. The interactive faceted search capability described is a demonstration how an approach based on these two technologies is more powerful and efficient for certain tasks as compared to traditional search engines.

The contributions of the chapter can be summarized as follows:

- novel indexing schema for semantic search, based on entity occurrence;
- description of a scalable implementation of such indexing;
- advanced faceted search interface, based on co-occurrence.

6.1 Introduction

The enormous scale of the web sets a new challenge – handling this scale, so that the users do not get lost and can efficiently perform their intended tasks. The ranking of documents and providing the most relevant information first is a step towards alleviating this information overload and has already been addressed to some extent by contemporary search engines. Another approach, which has not yet been seen in large scale applications, involves some linguistic and semantic analysis in order to better understand the user need and the data on the web and match
them at a deeper level. This analysis may result in the production of semantic
annotations, as defined in Kiryakov et al. (2005b), and represents an aspect of the
annotated resources that provides machine-readable insights into the meaning of
the content. Our belief is that the scalable and automatic extraction of semantic
annotations for the existing web is the route towards real-world Semantic Web
(SW) applications. Such metadata would allow SW applications to provide new
ways of searching, navigating, summarizing, and analyzing the web resources. For
several years, we have been developing the KIM Platform (Popov et al. 2004) for
semantic annotation, indexing and retrieval, which basic functionality is presented
in Sect. 6.1.1 below. In a nutshell, it allows for efficient semantic annotation and
indexing of documents with respect to named entities (NE). The basic indexing
approach is preceded by processing of the text content, which is then indexed with
a standard information retrieval (IR) engine, as discussed in Sect. 6.0. Such
indexing allows KIM to perform hybrid queries, combining full-text search,
istructured queries, and inference.

Recently we have extended KIM with a new module called CORE, focusing
on co-occurrence of entities and providing for incremental searching based on
this information, as well as ranking, tracking trends and popularity timelines of
these entities. Having some experience in extracting predefined relations
between named entities and understanding the extra investment of effort for
defining extraction algorithms for new relation types, we decided to extend KIM
in the direction of tracking associative (or soft) relations between entities, based
on their co-occurrence in the same block of content (or context). The information
gathered in the CORE module is based on the semantic annotations pro-
duced by KIM and provides another perspective to the results of the semantic
analysis of the content as compared to the semantic repository where all the
extracted entity instances are stored with their names, relations, and attributes.
CORE focuses on tracking some statistics for the bi-directional relations
between entities and documents, while leaving the semantically heavier descrip-
tions of entities to the semantic repository.

The KIM Platform is generally independent of the domain it is applied to, but
the concrete configuration context provided by default (ontology, instance base,
information extraction modules) is tuned for analysis of international news. As in
any domain, if one acquires large amounts of data to work on, interesting
co-relations of the domain entities could be monitored, shaping trends and depend-
encies that are not clear without the automatic assimilation of large volumes of
information. If the data is also aligned with particular points in time, timelines of
these co-relations or of the popularity of entities could be obtained.

The concrete solution presented here takes advantage of the specific strengths of
semantic repositories and the raw power of relational databases. The latter having
been developed over decades and capable of handling efficiently large volumes of
data with fixed structure (which is the case with the occurrence statistics); the
former allowing for inference and querying on top of formal knowledge. The com-
bination of the two in one system is like the unity of physical strength and prudent
mind in a person. The interactive faceted search (named CORE search) is a