Chapter 6
Approximate Knowledge Retrieval

6.1 Overview

The explosion of information on the World Wide Web and its moving to the next generation, so called Semantic Web (Berners-Lee, Hendler and Lassila 2001), is challenging the conventional relational database model at least in two aspects. Firstly, such information is no longer as simple as to be easily and naturally represented in relational tables. Secondly, exact answers are hardly obtained when searching for information on the Web.

For solving the first problem, fuzzy conceptual graphs have two advantages. Firstly, as a graph language, they are suitable for representation of weakly structured information on the Web. Secondly, in a graphical form, they are more flexible than fixed textbox patterns to describe queried objects in a system user interface. For readable Web data languages, after comparing the CG and the RDF, Berners-Lee (2001) concluded that the CG could be easily integrated with the Semantic Web. As shown later in details in Yao and Etzkorn (2004), there is a close mapping between the CG and the RDF, which has become a common and standard model for data interchange on the Web.

Meanwhile, the second problem is mainly concerned with the semantic distance between concepts, relations, names, and attributes in knowledge graphs, which has attracted significant research effort. For semantic distance between concept and relation types, previous methods can be classified into two main approaches, namely, corpus-based and ontology-based. In the former approach, the semantic distance between words representing concepts or relations is determined on the basis of their contexts of occurrence in a linguistic corpus. In the latter approach, it is determined on the basis of the semantic nets involving the words. However, such methods do not work for un-interpreted type labels in a hierarchy, because a word representing a type maybe just a dummy identifier for the type, not necessarily attached to any linguistic corpus nor having any meaning.

For matching knowledge graphs with concept and relation types defined in an ontology, Cao and Huynh (2006) proposed to use occurrence probabilities of entities and entity tuples of those concept and relation types based on their populations in the ontology of discourse. The rational of that proposal is that the semantic distance between two concepts depends on their populated instances in a particular context. For example, since in Viet Nam there are many motorcycles used, the semantic distance between the concept “motorcycle” is closer to the concept “vehicle” than their semantic distance in the USA. Cao and Huynh (2007) followed that population-based approach to define matching degrees for entity names represented by
character strings. Actually, it is often that, when querying, users do not remember or type in exactly names of entities to be searched for, which produces mismatching between a query and an answer. For fuzzy attribute values, fuzzy conditional probabilities were applied to measure their differences.

Another issue to be addressed is that, while symmetric measures like similarity have been discussed much in literature and commonly used in commercial search engines like Google, asymmetric measures appear to be less popular. In particular, in practice one may want to measure how much a query graph subsumes an answer graph. Applying fuzzy set theory to information retrieval, Miyamoto (1990) introduced the related term relation and the narrower term relation respectively as symmetric and unsymmetric ones. A short survey of soft information retrieval approaches was presented in Crestani and Pasi (1999). In Andreasen, Bulskov and Knappe (2003), although using the term “similarity”, which is presumably a symmetric measure, the authors argued that it should not be symmetric.

Regarding the system implementation aspect, realizing an approximate retrieval system of knowledge graphs from scratch is a formidable task. Meanwhile, there are new technologies for management of Web knowledge and information developed. For example, Sesame (Kampman, Harmelen and Broekstra 2002) with its query language SeRQL is for storing and querying RDF graphs. Those are open sources and have been used in many large-scale systems. They can be employed for approximate search using the query modification approach. That is, a query graph is first generalized to retrieve stored knowledge graphs by the exact matching mechanism of those tools. The retrieved graphs, which are subsumed by the generalized query graph, are then matched to the original query graph to give their similarity and subsumption degrees.

Since a formal language like SeRQL, with various syntactic symbols for machinery processing, is still difficult for non-experts to use, in this chapter we present conceptual graphs as an alternative graph-based language for query expressions at the interface layer. Query conceptual graphs are then automatically converted to SeRQL queries to be executed by the Sesame query engine.

The rest of the chapter is organized as follows. Section 6.2 reviews previous definitions of similarity measures for words, and proposes ontology-based similarity and subsumption measures for entity concept types, relation types, names, and fuzzy attribute values. Section 6.3 presents the use of Sesame, SeRQL, and conceptual graphs for storing and querying knowledge graphs. Section 6.4 develops a query modification method and a CG-SeRQL mapping algorithm for approximate knowledge graph retrieval. Section 6.5 presents application of the defined measures and methods to development of our knowledge and information management system VN-KIM. Finally, Section 6.6 concludes the chapter.

6.2 Matching Measures for Entity Types, Names, and Attributes

Similarity between Words

The similarity between expressions, such as words or concepts, measures how close their semantics or meanings are. Previous methods for words could be