Chapter 69
Semantic Search Exploiting Formal Concept Analysis, Rough Sets, and Wikipedia

Yuncheng Jiang
South China Normal University, China

Mingxuan Yang
South China Normal University, China

ABSTRACT

This article describes how the traditional web search is essentially based on a combination of textual keyword searches with an importance ranking of the documents depending on the link structure of the web. However, one of the dimensions that has not been captured to its full extent is that of semantics. Currently, combining search and semantics gives birth to the idea of the semantic search. The purpose of this article is to present some new methods to semantic search to solve some shortcomings of existing approaches. Concretely, the authors propose two novel methods to semantic search by combining formal concept analysis, rough set theory, and similarity reasoning. In particular, the authors use Wikipedia to compute the similarity of concepts (i.e., keywords). The experimental results show that the authors' proposals perform better than some of the most representative similarity search methods and sustain the intuitions with respect to human judgements.

1. INTRODUCTION

The World Wide Web is the world's most valuable information resource and has become the world's largest database with search being the main tool that enables organisations and individuals to exploit its huge amounts of information that is freely offering (Virgilio et al., 2012). Thus, having a successful mechanism for finding and retrieving the most relevant information to a task at hand is of major importance. Traditionally, Web search is essentially based on a combination of textual keyword search with an importance ranking of the documents depending on the link structure of the Web (Fazzinga and

DOI: 10.4018/978-1-7998-0951-7.ch069

Lukasiewicz 2010). However, one of the dimensions that has not been captured to its full extent is that of semantics. Combining search and semantics gives birth to the idea of the semantic search. Semantic search can be described in a sentence as the effort of improving the accuracy of the search process by understanding the context and limiting the ambiguity (Melo et al., 2016; Virgilio et al., 2012; Vocht et al., 2017). In fact, the issue of semantics has become the grand challenge for the next-generation World Wide Web (Jindal et al., 2014; Storey et al., 2008).

The development of semantic search is currently an extremely hot topic (Binding et al. 2015; Fazzinga and Lukasiewicz, 2010; Formica, 2012; Formica et al., 2013; Likavec et al., 2015; Melo et al., 2016; Storey et al., 2008; Virgilio et al., 2012; Vocht et al., 2017). For example, Formica (2012) showed how rough set theory (Pawlak 1991) could be employed in combination with fuzzy formal concept analysis to perform Semantic Web search and discovery of information in the web. Formica et al. (2013) pointed out that there were many proposals on semantic search and retrieval in the literature, but there was still no specific solution that clearly emerged. For this reason, they presented a proposal based on a notion that is currently gaining momentum in the field: the Information Content (IC) approach (Resnik, 1995; Jiang et al., 2017).

However, there are still some limitations in the above approaches. For example, similarity reasoning in (Formica et al., 2013) was based on ontology based IC computation (Sanchez et al., 2011). The fact that ontology based IC computation only relies on ontological knowledge is a drawback because it completely depends on the degree of coverage and detail of the unique input ontology (Sanchez and Batet, 2013). Especially, with the emergence of social networks (Martinez-Gil and Aldana-Montes 2013), a lot of (sets of) concepts or terms are not included in domain ontologies (Kumar et al., 2017), therefore, IC computation that is based on these kinds of domain ontologies cannot be used in these tasks. On the other hand, the prerequisite of ontology based IC computation is the existence of several predefined domain ontologies. Clearly, the construction of domain ontologies is time-consuming and error-prone and maintaining these ontologies also requires a lot of effort from experts. Thus, the methods of ontology based IC computation are also limited in scope and scalability (Jiang et al., 2015). These limitations of similarity reasoning are the motivation behind the new techniques presented in this paper which implement semantic search by exploiting Wikipedia (Hovy et al., 2013; Medelyan et al., 2009). It should be noted that the authors' approaches are only suitable for less specific domains because Wikipedia does not guarantee the coverage for specific domains. In the case of a very specific domain, a domain specific reference resource such as ontologies is needed. On the other hand, in the case of less specific domains, semantic search relies on some knowledge sources such as WordNet (Fellbaum, 1998), DBpedia (Lehmann et al., 2009; Li et al., 2017b), Wikipedia (Hovy et al., 2013; Medelyan et al., 2009), and YAGO (Hoffart et al., 2013; Suchanek et al., 2008). In this paper, the authors will exploit Wikipedia to implement semantic search. In fact, the authors' approaches can also be based on other knowledge sources such as DBpedia or YAGO by simple adjustment.

The purpose of this paper is to present some new methods to semantic search to solve the shortcomings of existing approaches. Concretely, along the line of the research of (Formica 2012; Formica et al. 2013), the authors will propose some novel methods to semantic search by combining formal concept analysis (Ganter and Wille, 1999), rough set theory (Pawlak, 1991), and Wikipedia based similarity reasoning (Jiang et al., 2017; Li et al., 2017a; Ponzetto and Strube, 2007; Taieb et al., 2013; Yazdani and Popescu-Belis, 2013). On the other hand, semantic search usually adopts user-friendly unstructured (Rocha et al., 2004) or semi-structured (Anyanwu et al., 2005) query strategies. However, these reported approaches are still much less attractive than keyword search, which is the most effective and successful paradigm