Web Semantics: Science, Services and Agents on the World Wide Web 29 (2014) 39–52

Contents lists available at ScienceDirect

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

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

A hybrid approach to finding relevant social media content for complex domain specific information needs

Delroy Cameron *, Amit P. Sheth, Nishita Jaykumar, Krishnaprasad Thirunarayan, Gaurish Anand, Gary A. Smith

Ohio Center of Excellence in Knowledge-enabled Computing (Kno.e.sis), Wright State University, Dayton OH 45435, USA

HIGHLIGHTS

- Complex domain specific search requires richer models to be more effective.
- Structured knowledge and intelligible constructs together can enhance query interpretation.
- A hybrid search platform is more effective for practical domain specific information needs.

ARTICLE INFO

Article history:
Received 1 August 2013
Received in revised form
7 September 2014
Accepted 2 November 2014
Available online 15 November 2014

Keywords:
Semantic search
Domain specific information retrieval
Complex information needs
Ontology
Background knowledge
Context-free grammar
Knowledge-aware search

ABSTRACT

While contemporary semantic search systems offer to improve classical keyword-based search, they are not always adequate for complex domain specific information needs. The domain of prescription drug abuse, for example, requires knowledge of both ontological concepts and "intelligible constructs" not typically modeled in ontologies. These intelligible constructs convey essential information that include notions of intensity, frequency, interval, dosage, and sentiments, which could be important to the holistic needs of the information seeker. In this paper, we present a hybrid approach to domain specific information retrieval (or knowledge-aware search) that integrates ontology-driven query interpretation with synonym-based query expansion, and domain specific rules, to facilitate search. Our framework is based on a context-free grammar (CFG) that defines the query language of constructs interpretable by the search system. The grammar provides two levels of semantic interpretation: (1) a top-level CFG that facilitates retrieval of diverse textual patterns, which belong to broad templates and (2) a low-level CFG that enables interpretation of specific expressions that belong to such patterns. These low-level expressions occur as concepts from four different categories of data: (1) ontological concepts, (2) concepts in lexicons (such as emotions and sentiments), (3) concepts in lexicons with only partial ontology representation, called *lexico-ontology* concepts (such as side effects and routes of administration (ROA)), and (4) domain specific expressions (such as date, time, interval, frequency, and dosage) derived solely through rules. Our approach is embodied in a novel Semantic Web platform called PREDOSE, which provides search support for complex domain specific information needs in prescription drug abuse epidemiology. When applied to a corpus of over 1 million drug abuse-related web forum posts, our search framework proved effective in retrieving relevant documents when compared with three existing search systems.

Published by Elsevier B.V.

1. Introduction

The use of structured background knowledge (ontologies) to enhance search has gained considerable traction among contemporary information retrieval systems. Ontologies offer to improve search by capturing the meaning of real-world concepts and their associations. The formal representations modeled in ontologies have been used to positively impact many complex tasks, including interoperability, personalization, and knowledge discovery.

While semantic search has gained credibility, compared to classical keyword-based and hyperlinked-based search, there is often a misalignment between the information needs of users and the knowledge model developed to meet such needs. Ontologies provide a means for interpreting some elements of complex information needs, but not all aspects of such needs [1]. The main issue

* Corresponding author. Tel.: +1 937 775 5213; fax: +1 937 775 5133.
E-mail address: delroy@knoesis.org (D. Cameron).

http://dx.doi.org/10.1016/j.websem.2014.11.002
1570-8268/Published by Elsevier B.V.

is that ontologies often have limited scope, while users are unrestricted in the range of information they can seek on a given topic. A user information need can transcend data types and sources, exceeding what is formally modeled.

In spite of this, many semantic search applications [2–5], semantic search engines (Hakia, Bing), and hybrid information retrieval approaches [1,6–8] rely heavily on ontologies for query interpretation. While these approaches serve their intended purpose, they are generally unsuitable for domain specific applications, such as prescription drug abuse. General-purpose search engines such as Google and Yahoo that rely on keyword-based and hyperlinked-based models, may not perform well on domain specific data. This is because minimal (and often inadequate) support is provided for interpreting the additional elements that could be important to an information need, but not formally captured by the knowledge model.

We address this problem by developing and evaluating a hybrid approach to search (or knowledge-aware search) that allows query specification and interpretation of diverse expressions, which are involved in various aspects of complex information needs. To illustrate our approach, consider a scenario in which an epidemiologist in the domain of prescription drug abuse is seeking insights into emerging patterns and trends in drug abuse using social media. For brevity, we present only one of many information needs explored in PREDOSE (http://wiki.knoesis.org/index.php/PREDOSE).

Information need: How are drug users engaging in the use of the semi-synthetic opioid Buprenorphine, through excessive daily dosage?

Inherent in this information need is the following relevant background knowledge. Buprenorphine is an opioid antagonist used in the treatment of opioid addiction, including addiction to Heroin, OxyContin, and Vicodin. Prescribed daily dosage varies by individual ranging from 4–32 mg.[1] Buprenorphine is known to stabilize drug users from withdrawal symptoms, but can also induce an opiated effect. This treatment drug is therefore at risk for abuse. Epidemiologists are interested in understanding the dosage practices of Buprenorphine users, including amounts taken, frequency of use, and side effect experienced, to better understand emerging patterns and trends of abuse.

A suitable user query provided by a domain expert may involve the following keywords: "buprenorphine dosage exceed 4 mg daily". A robust search system may correctly interpret the keyword 'buprenorphine' as the standard DBpedia resource: http://dbpedia.org/resource/Buprenorphine. Then through non-trivial query expansion, the system may also associate the keywords 'bupe', 'bupey', 'suboxone', 'subbies' and 'suboxone film', with 'Buprenorphine', as synonyms. Likewise, the search system may expand the keyword 'daily' with the synonyms 'day', 'night', 'morning', and 'afternoon', using available (or manually created) lexicons that contain such mappings. However, the intricate challenge is interpreting the notion of excessive dosage, expressed as the phrase "dosage exceed 4 mg".

In the development of Active Semantic Electronic Medical Records (ASEMR), Sheth et al. [9] created rules expressed in RDQL [10] (precursor to SPARQL) to enable specification of additional constructs (including dosage) that compensate for deficiencies in the knowledge model. Similarly, in the Semantic Content Organization and Retrieval Engine (SCORE) [11,12], Hammond et al. implemented various rules derived using regular expressions to specify quantity-conveying metadata (such as 'currency', 'percentage', 'amount', 'time' and 'dates') which were not present in the ontology. In the Knowledge and Information Management platform (KIM) [13], Popov et al. modeled various lexical resources in the ontology, such as currency, dates and abbreviations, which were subsequently used for document annotation. However, the information need presented here requires a more in-depth interpretation.

To appropriately interpret excessive dosage, the notion of dosage itself must first be specified using its constituent members: DOSAGE-OPERATOR (e.g., '>', '<'), DOSAGE-AMOUNT (e.g., '4', '10'), and DOSAGE-UNIT (e.g., 'mg', 'tablet'). In this way, the search term '>4 mg' could be an abstraction of the search phrase "dosage exceed 4 mg". Rules must then be used to interpret each constituent according to what is possible in the corpus. This is important because a DOSAGE-UNIT may have various lexical representations in text (e.g., mg, milligram, milli-gram). Likewise, the DOSAGE-OPERATOR can have multiple equivalent manifestations (such as '>', 'greater than', 'more than' and 'above'). Similarly, DOSAGE-AMOUNT can be numeric or textual. According to these possible interpretations, '6 mg', 'ten milligrams', 'about 8 mg', 'a bit more than 30 mg' etc., are all valid expressions for the query 'dosage exceed 4 mg.' The matching documents for the entire query ("buprenorphine dosage exceed 4mg daily") are obtained after filtering heuristics are applied to retrieve text fragments from the corpus that match the interpretation of each query component. In this way, a hybrid approach to information retrieval would have been utilized, which leverages ontologies, lexicons, and rules for query interpretation of domain specific data.

Concretely, our approach is based on a context-free grammar (CFG) that defines the query language of constructs interpretable by the search system. The grammar provides two levels of semantic interpretation: (1) a top-level CFG defines broad patterns that can be interpreted by the system and (2) a low-level CFG defines the interpretation of certain specific elements that belong to such patterns. The query language of the grammar is specified in an IBM declarative information extraction specification called SystemT [14,15], which is designed for information extraction from heterogeneous texts. SystemT is advantageous because it enables porting of rules to texts in other domains. Some of these domains specifically: (1) biomaterials and materials science, (2) cannabis and synthetic cannabinoid research, and (3) clinical texts on cardiology reports.

In an evaluation using a corpus of over 1 million web forum posts related to drug abuse, our hybrid search system retrieved a larger number of relevant documents when compared with three existing search systems. These systems are the: (1) semantic search engine Hakia, (2) crowd sourcing-based search engine DuckDuckGo, and (3) popular search engine Google. Note that since these search engines are not specifically engineered to handle domain specific data, our results are not surprising. However, our experiments highlight the need for more effective approaches to domain specific search as noted in [16]. The specific contributions of this research are as follows:

- We develop a hybrid approach to domain specific information retrieval that interprets four categories of data. These are: (1) structured background knowledge in ontologies, (2) concepts in lexicons; (3) concepts in lexicons with partial ontology representation called *lexico-ontology* concepts (see Section 2.1.2), and (4) concepts defined using rules.
- We utilize a CFG to formally define the query language of strings interpretable by the system. The CFG provides two levels of semantic interpretation: (1) a top-level CFG for interpreting general textual patterns and (2) a low-level CFG for interpreting specific expressions.
- We show that our approach is effective through an evaluation against three popular search systems.

[1] Note that the actual amounts used in examples throughout this manuscript are anecdotal only.