Science of Computer Programming 94 (2014) 32–52

Contents lists available at ScienceDirect

Science of Computer Programming

www.elsevier.com/locate/scico

A query language and ranking algorithm for news items in the Hermes news processing framework

Frederik Hogenboom *, Damir Vandic, Flavius Frasincar, Arnout Verheij, Allard Kleijn

Erasmus University Rotterdam, P.O. Box 1738, NL-3000 DR, Rotterdam, The Netherlands

HIGHLIGHTS

- We describe a graphical query language for news, HGQL.
- HGQL supports the conjunction, disjunction, and negation operators.
- We develop an extended Boolean model ranking algorithm with negation support.
- We show that our ranking algorithm outperforms three other ranking algorithms.
- We show that HGQL outperforms its text-based counterpart.

ARTICLE INFO

Article history:
Received 16 October 2012
Received in revised form 29 May 2013
Accepted 30 July 2013
Available online 13 August 2013

Keywords:
Query languages
News ranking
Ontology-based querying

ABSTRACT

Hermes is a Web-based framework that makes use of many Semantic Web technologies for building personalized news services. Ontologies are employed for knowledge representation, natural language processing techniques are used for semantic text analysis, and semantic query languages enable the specification of the desired information. To accommodate for the need for an intuitive way to create complex queries for news information, we present the Hermes Graphical Query Language (HGQL). The language enables users to create structured queries that use disjunctive, conjunctive, negation, and pattern operators. In addition, this paper presents a ranking algorithm based on the queries made using our graphical query language. Results show that our proposed ranking algorithm significantly outperforms three state-of-the-art ranking algorithms and that users prefer our graphical query language over a text-based alternative.

1. Introduction

One of the major problems that arise as a result of today's unstoppable growth of the Web, is the information overload daily users are confronted with. This calls for a methodical approach to information filtering, for example news recommendation [1–3], in such a way that the presented subset of results actually represents the individual user's preferences. Many information retrieval techniques are already available, of which keyword matching is the most common one. In such approaches, user-specified keywords are matched to the available textual data, resulting in a relevant selection of information matching these keywords. However, a problem with this approach is the lack of semantics, i.e., the meaning of words is not taken into account. For example, the keyword 'apple' could refer to fruit, the company, or even a person's name.

* Corresponding author. Tel.: +31 (0)10 408 8907; fax: +31 (0)10 408 9031.
E-mail addresses: fhogenboom@ese.eur.nl (F. Hogenboom), vandic@ese.eur.nl (D. Vandic), frasincar@ese.eur.nl (F. Frasincar), 308057av@student.eur.nl (A. Verheij), 303118ak@student.eur.nl (A. Kleijn).

0167-6423/$ – see front matter
http://dx.doi.org/10.1016/j.scico.2013.07.018

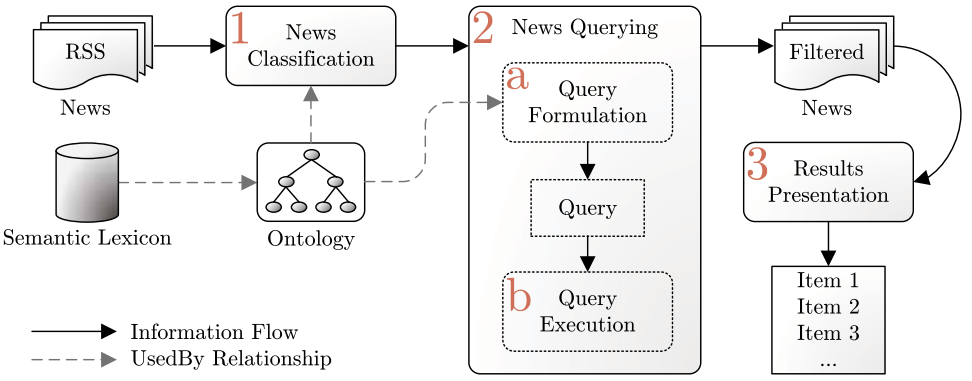

Fig. 1. Overview of the Hermes framework.

In order to deal with the previously identified semantic issue while processing (news) text, in earlier work we have introduced the Hermes news personalization framework [4,5] which is built upon Semantic Web technologies. The news items are gathered from RSS feeds provided by the user. Hermes is composed of multiple natural language processing resources that enable the processing (and querying) of news. Also, by employing a plug-in-based software architecture, the Hermes framework allows for the addition of user-created specialized information processing plug-ins. By default, the framework stores lexicalized domain concepts and relations (i.e., properties that relate concepts to each other or to data types) in a domain ontology. The ontology also stores synonyms (string representations) of domain-specific entities like companies, persons, etc., as well as their relations, such as subsidiary and competitor relations. The domain ontology is used to index news as well as to retrieve relevant news items in a semantically-enhanced way. In addition, we have proposed ontology-based recommendation plug-ins that also benefit from Hermes' ontology and the news processing framework [2,6].

However, semantics-based matching alone is not enough in order to provide a good personalized news service, as user preferences also need to be elicited. This could be achieved by letting the user specify queries that express the concepts of interest. Therefore, in previous work we have devised a text-based query language that makes use of linguistic patterns that incorporate lexical, syntactic, and semantic elements [7], and we have successfully implemented the language into Hermes as a plug-in. However, in order to aid the average, day-to-day user, who is not an expert in information technology, with creating a query, graphical query languages can prove to be useful here. These structured graphical languages aim to minimize the amount of effort the user has to put into formulating queries, by providing a less complex syntax, and by using only Boolean (AND, OR, and NOT) and sequence operators (in the form of a sentence, i.e., they have a subject, a predicate, and an object). Based on our previous experience with creating a graphical query language for RDF, RDF-GL [8], in recent work [9], we have introduced a graphical query language designed for Hermes: the Hermes Graphical Query Language (HGQL).

Although semantics-based queries generated through a graphical query language provide the user with a subset of potentially interesting news items, these items need to be ranked according to their relevance to the user query. Several weighting schemes for concept importance have been proposed in the literature. Most of the schemes can cope with AND and/or OR operators for queries, but few solutions have been devised to use these operators together with the NOT operator. Hence, we propose to enhance the extended Boolean model [10] with the negation operator.

This paper builds on recent work [9] and has four main contributions with respect to the state-of-the-art. First, we describe a graphical query language for searching news that goes beyond current keyword-based approaches, and which is additionally supported by an implementation in Hermes. Compared to our earlier work, we provide more details on the main language elements and grammar, as well as the Hermes framework and the implementation of HGQL. Second, we devise a ranking algorithm for sorting news that effectively supports the negation operator. Third, in contrast to our recent work, in our current endeavors, we provide a quantitative and qualitative evaluation of the proposed query language. Last, we provide an extensive, more detailed, evaluation of our ranking algorithm with respect to several vector space model weighting schemes.

The rest of the paper is organized as follows. First, Section 2 provides a more thorough description of the Hermes framework. Next, Section 3 describes related work on graphical query languages and relevance ranking algorithms. Section 4 proposes the Hermes Graphical Query Language (HGQL). Section 5 devises a ranking algorithm for HGQL, and Section 6 discusses the implementation of the language and its ranking algorithm. The query language and algorithm are evaluated against other approaches in Section 7. Last, Section 8 presents our conclusions and suggests future work.

2. Hermes

The Hermes framework [1,4], as depicted in Fig. 1, is comprised of a sequence of steps for building a personalized news service. The system's inputs are RSS news feeds, whereas its outputs are filtered (relevant) news items. The core of the Hermes framework is a domain ontology developed by domain experts, employed for indexing news items and for