Expert Systems With Applications 66 (2016) 189–202

Contents lists available at ScienceDirect

Expert Systems With Applications

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

Upgrade of a current research information system with ontologically supported semantic search engine

Aleš Bošnjak[a], Vili Podgorelec[b,*]

[a] Institute of Information Science, Prešernova 17, SI-2000 Maribor, Slovenia
[b] University of Maribor, FERI, Institute of Informatics, Smetanova 17, SI-2000 Maribor, Slovenia

ARTICLE INFO

Article history:
Received 20 April 2016
Revised 2 September 2016
Accepted 7 September 2016
Available online 10 September 2016

Keywords:
Current research information systems
Metadata models for ontology
Ontological infrastructure
Semantic search
Faceted information search

ABSTRACT

Harmonising the metadata format alone does not solve the issue of efficient access to relevant information in heterogeneous environments, when different systems use different content, contextual and semantic concepts for certain entities. One such type of heterogeneous systems are also Current Research Information Systems (CRIS), which store their data primarily in local relational databases, using different formats and various local concepts.

In this article, we study the possibilities and propose a new ontologically supported semantic search engine (OSSSE) which, in addition to the harmonisation of the metadata format among local CRIS systems, also ensures that the meaning of data and/or concepts that belong to various metadata entities are also harmonised. A special model of ontological infrastructure was designed, and dedicated test ontology was created alongside with a new simplified algorithm for creating ontology, the basis of which is the distinction between new and already existing classes in terms of content. Finally, we evaluated the proposed OSSSE model using a simulation of the search process on the base of 41,113 real searches within SICRIS. The obtained results show that regardless of the search situation, the proposed OSSSE is always at least as efficient as a search without ontological support in terms of precision, while recall remains the same; the improvement has been shown to be statistically significant with a high confidence interval ($p<0.005$).

The proposed OSSSE model is able to solve the issue of harmonizing the data where different heterogeneous systems use different content, contextual and semantic concepts, which is the case in many advanced expert systems. In this manner, the more the search is carried out based on the properties described by the supporting ontology, the more the infrastructure can help a searcher. The proposed concepts, ontological infrastructure and the designed semantic search engine may well help to improve search precision in several information retrieval systems.

1. Introduction

The issue of efficient access to relevant information in heterogeneous environments and distributed complex information systems is well-known and is usually solved with approaches involving federated search modes (Jacsó, 2004) or with search modes including cache (Graham, Jarnevich, Simpson, Newman, & Stohlgren, 2011). One type of heterogeneous systems are also Current Research Information Systems, known by the English acronym CRIS in Europe (Jeffery & Asserson, 2006). Because most European CRIS systems store their data in relational databases and the systems are heterogeneous, addressing the issue of efficient access to information in CRIS systems also represents the actual address of the wider problem of access to information, stored in heterogeneous relational databases.

Even though the European umbrella organisation EUROCRIS recommends the metadata format CERIF (EUROCRIS, 2008), the European formats partially differ when applied due to historic and other reasons. To support interoperability between individual CRIS systems, the CERIF Exchange metadata format has been mutually agreed on (Joerg, Krast, Jeffery, & Van Grootel, 2007) and must be supported by all local CRIS systems that wish to cooperate with each other.

On some characteristic examples taken from the Slovenian CRIS system (whereas the same applies to most other local CRIS systems) it can be demonstrated that harmonising the metadata format is not enough as different CRIS systems use different content,

* Corresponding author.
E-mail addresses: ales.bosnjak@izum.si (A. Bošnjak), vili.podgorelec@um.si (V. Podgorelec).

http://dx.doi.org/10.1016/j.eswa.2016.09.011
0957-4174/

contextual and semantic concepts for certain entities. For example: the concept of a researcher that is registered with a national research agency and is as such eligible to apply for national projects and programmes must comply with different rules in Germany than in Slovenia. For this purpose, in addition to the harmonisation of the metadata format among local CRIS systems, it must also be ensured that the meaning of data and/or concepts that belong to various metadata entities are also harmonised.

A potential solution to the issues described above could be the use of a semantically enriched search engine supported by the adequate ontology. In general, information content can be simply represented and the changes maintained with the use of ontologies (Berners-Lee, Hendler, & Lassila, 2001) that are, as the basis of semantic solutions, one of the most widely known and tested approaches to formalised representation of the meaning of data and information. On the other hand, interoperability should be ensured via a common search engine for all local CRIS systems.

Semantically enriched search engines can be structurally divided into two larger groups:

(a) search engines where the search is performed with one single semantically enriched query (Sakthi Murugan, Shanthi Bala, & Aghila, 2013), and
(b) search engines where the search is performed in several queries; the first query is for ontology, the second for all individual databases (Chen, Li, & Wang, 2009).

For the purpose of developing a semantically enriched search engine in our case, the second group is more suitable. In this manner, we developed a prototype of a semantically enriched common search engine for local CRIS systems, based on the Ontobroker ontological infrastructure (Ontobroker, 2016), which is presented in this paper. With the use of a separate inference engine, the consistency of the ontology is checked dynamically with every change, while it also ensures efficient separation of basic ontological concepts from temporary ones, generated from the basic concepts and their properties throughout the inference process during the use of the semantic search engine (SSE). The possibility of control and separation between basic and generated ontology concepts are of great importance for the production of this ontological infrastructure, as these concepts produce new metadata, crucial for the improvement of the search engine with the implementation of a faceted search mode.

A faceted search mode for narrowing down search results uses facets that are usually created in systems without ontological support from the contents of those metadata entities with controlled entry (Allen, 2005). In our developed ontologically supported SSE, ontologically acquired additional metadata with its contents is added to the facets. Because the additional metadata, which is ontologically acquired both from the basic concepts and the concepts generated through the process of inference, is conceptually linked to the search metadata, it is very relevant to the existing search as additional facets, which is not generally the case for other standard facets. The conceptual relationship is guaranteed through the use of semantic query based on descriptive logic (DL) (Pérez-Urbina, Horrocks, & Motik, 2009).

To evaluate the influence of ontologically acquired metadata on search results, the impact of ontological infrastructure on precision and recall of obtained search results was calculated based on a set of characteristic examples. The results indicate that the improved quantity of additional metadata and, as a result, also the search precision, depend on how many concepts are included in the ontology and on whether these concepts coincide with the search request. In cases where the additional metadata corresponds to the search requests, narrowing down increases the ratio of relevant search results and the precision increases. If the additional metadata does not correspond with the search request, precision remains the same. Recall remains the same in both cases.

In addition to the introduced model of the ontological infrastructure of a SSE, the paper also presents an optimised procedure for creating a support ontology where only those local concepts are included in the ontology that create new classes in terms of content. Local concepts are those that are only applicable to an individual local CRIS system (e.g. Slovenian SICRIS, Dutch METIS, or Norwegian CRISTIN) and do not necessarily apply to all CRIS systems on a global level. In the process of building an ontology, only those classes are entered that are not yet included in the ontology and do not have an equivalent. In the case of the used test ontology, such an optimisation presents approximately 40% fewer recorded local concepts. In general, we estimate that for larger ontologies (because of a more frequent appearance of inheritance and interconnection between classes) a higher optimisation degree is expected, so we think that such a rationalisation in the process of building a common European integrated CRIS ontology, where a large quantity of local concepts is included, would be more than welcome.

The rest of the paper is organised as follows: Section 2 presents the model of European CRIS systems and the recommended way of aggregating them via the CERIF Exchange format as well as the disadvantages and problems arising from such a model. Section 3 describes in great detail the suggested upgrade with an ontologically supported semantic search engine (OSSSE) for CRIS systems that represents a possible solution to the identified problems of access to relevant information in a heterogeneous environment of various local CRIS systems. Section 4 deals with the evaluation of the proposed OSSSE. Section 5 is dedicated to the discussion about the influence of the suggested model to the solution of the larger issue of combining several information systems that store their data in heterogeneous relational databases. Section 6 summarises the main findings and a few guidelines for future work are given.

2. Heterogeneous European CRIS systems and an attempt of ensuring interoperability between them

The Current Research Information System (CRIS) is intended for the informatization of managing the research process. EUROCRIS is in charge of European CRIS systems; they are a non-profit umbrella organisation where individual administrators and managers of European current research information systems have been uniting their professional interests for over a decade (EUROCRIS, 2016). Since its foundation, the organisation has been working towards connecting the CRIS systems of individual countries into one common European system in the most effective way.

The CERIF format (EUROCRIS, 2008) is a standard in its application, although it technically has the status of an EU recommendation. The EUROCRIS organisation is in charge of CERIF (EUROCRIS, 2016). In its current development phase it is called CERIF2008 – 1.2 Full Data Model (Joerg, Jeffery, Van Grootel, Asserson, & Dvorak). It is presented in the form of an entity-relationship (ER) model in Fig. 1.

CERIF is adapted to relational databases that describe the properties of the basic research entities (indicated in green in Fig. 1) in their basic tables:

- `Person` – to describe persons included in the research process
- `OrganisationUnit` – to describe research organisations or research units
- `Project` – to describe research projects or programmes

Additionally, the model contains three result entities (indicated in orange in Fig. 1):