Search Ontology, a new approach towards Semantic Search

Alexandr Uciteli[1] Christoph Goller[2] Patryk Burek[1]
Sebastian Siemoleit[1] Breno Faria[2] Halyna Galanzina[2] Timo Weiland[3]
Doreen Drechsler-Hake[3] Wolfram Bartussek[2,3]
Heinrich Herre[1]

[1]Institute for Medical Informatics, Statistics and Epidemiology (IMISE)
Leipzig University
alexander.uciteli@imise.uni-leipzig.de

[2]IntraFind Software AG [3]novineon CRO & Consulting Ltd
christoph.goller@intrafind.de

Abstract:
We present an innovative system for semantic search, based on an ontology, called Search Ontology. The Search Ontology contains search terms, for which synonymous labels can be defined, and search concepts, specified by rules, that determine how search terms are combined with abstract NEAR or Boolean operators to describe corresponding concepts in documents. A search query can be generated from the ontological specification and executed on an information retrieval system such as Lucene[1] afterwards. This approach has the advantage that the user can create powerful and complex queries by ontological specifications only, with minimal effort and without knowing the query syntax. The ontology itself is easily adaptable, extensible and reusable. No information contained in the ontology is used while preprocessing and indexing the documents, since the ontology is being constantly expanded by users of the system and changes in the ontology should not trigger new indexing and analysis for the whole document collection. The system is intended for domain experts, e.g., patent examiners or experts in the field of post-market surveillance of medical devices.

1 Introduction

The need to retraceably gather documents from heterogeneous sources in order to substantiate answers to very specific questions is a recurring scenario in many domains, e.g., patent examination and medical post-market surveillance (PMS). The quality of the search results depends on several aspects, including the possibility to specify powerful queries. The central component of the expounded search method is an ontology, called Search Ontology (SO), that allows the expert to formally specify domain concepts, search terms associated to the domain, and rules describing domain concepts. Furthermore, a prototypical software tool has been developed which generates Lucene queries from information contained in the ontology. These queries can be used by a suitable search engine (Solr[2],

[1] http://lucene.apache.org/
[2] http://lucene.apache.org/solr/

Elasticsearch[3]). This method supports an autonomous construction of complex queries by a domain expert and allows an easy adaptability of the ontology and the reusability of its parts for recurring tasks.

The ideas, presented in this paper[4], have been developed within the BMBF[5]-supported project OntoVigilance[6]. The aim of this project is the development of a tool which supports a systematic gathering of clinical safety data of medical devices. Each manufacturer is obliged to set up a system of PMS. The primary purpose is protection of the patients safety and development of safer medical devices for public health. However, the retrieval of suitable information for a rule-consistent PMS is hampered by highly heterogeneous data sources and a whole amount of data sets to be evaluated.

2 Detailed Description of the Approach

2.1 Search Ontology (SO)

The SO is developed in Web Ontology Language (OWL)[7]. The categories of the ontology (see **Categories**) are represented in OWL by classes, the relations (see **Relations**) - by properties and the rules (see **Rules**) - by property restrictions. The SO can be used for information retrieval in any domain. For that, it must be extended by the corresponding domain ontology. In the OntoVigilance project an ontology for the post-market surveillance and vigilance domain (PVO) is developed. The PVO enables the identification of available scientific data on safety, complications and efficacy of a medical device in the field of endoscopy. For example, recurrence of bleeding (re-bleeding) after an unsuccessful endoscopic intervention due to device failure or user errors is a life threatening complication. A faulty designed endoscopic tool, an improper handling or unsuitable anatomical location for the endoscopic device may lead to failure in hemostasis. By identifying failure modes and root causes of device failures and user errors, the design or instructions for use given for such medical devices may be improved to further prevent such complications.

We use the top level ontology GFO (General Formal Ontology, [Her10]) as a framework for the foundation of the Search Ontologys components. GFO provides, among others, an ontology of categories. The instances of the category GFO:Concept are concepts, e.g., the concept Hemostasis, whereas the instances of the category GFO:Symbolic_Structure are abstract strings, e.g., the abstract string "Hemostasis", the instances of which are tokens, written, e.g., on a sheet of paper.

[3] http://www.elasticsearch.org/

[4] We present work in progress. Only parts of the ideas described are already implemented. The ontologies and tools will be licensed after completion of the project.

[5] Federal Ministry of Education and Research; http://www.bmbf.de

[6] http://ontovigilance.org/

[7] http://www.w3.org/TR/owl2-overview/

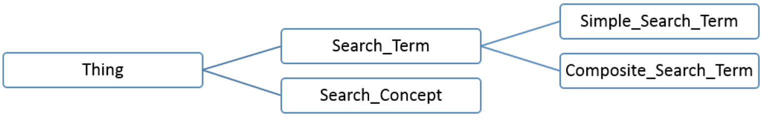

Figure 1: Categories of the Search Ontology

Categories The SO includes - on the top level - two categories, called Search_Term and Search_Concept. The category Search_Concept is a subcategory of GFO:Concept and the category Search_Term is a subcategory of GFO:Symbolic_Structure. The category Search_Term has two subcategories: Simple_Search_Term and Composite_Search_Term (see Fig. 1). The category Simple_Search_Term is instantiated by single words, while the instances of Composite_Search_Term are composed of the instances of Simple_Search_Term by using a relation has_part (see **Relations**), according to certain rules (see **Rules**). The domain experts may define various subcategories of Simple_Search_Term resp. Composite_Search_Term which can be used to structure the search terms of the domain. Examples of subcategories of Simple_Search_Term from the PVO (see Fig. 2) are Hemostasis_Adjective (having instances: inadequate, incomplete), Hemostasis_Noun (with instance: hemostasis) and Rebleeding_Noun (with instance: rebleeding). An example of a subcategory of Composite_Search_Term is Rebleeding_Phrase. The instances of this category are combinations of an instance of Hemostasis_Adjective and an instance of Hemostasis_Noun (e.g., “inadequate hemostasis”, “incomplete hemostasis”). To any instance of the category Simple_Search_Term may be associated various labels. By such an association terms may be denoted by several synonymous notations resp. notations in different languages. The category Search_Concept is the root node of the subtree of domain concepts. For any domain concept a search rule can be specified (see **Rules**). Examples of concepts of the PVO are Complication or its subcategory Rebleeding.

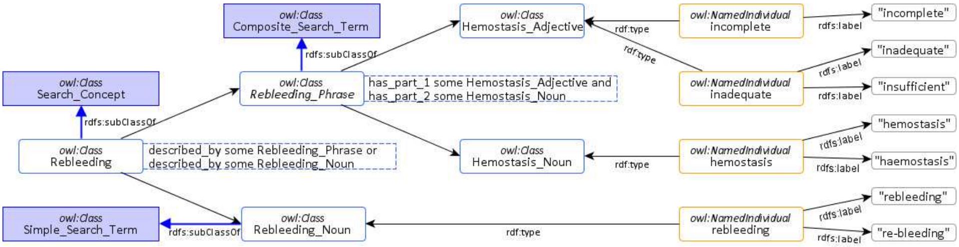

Figure 2: Example

Relations The SO includes three relations: has_part, max_distance and described_by. The first two relations connect instances of Search_Term, whereas the relation described_by connects instances of Search_Concept with instances of Search_Term. By using the relation has_part instances of Composite_Search_Term can be assembled from instances of Simple_Search_Term. The relation max_distance describes the maximal number of words which are allowed to occur between the simple terms within the composite term. The relation described_by specifies a connection between the domain concepts and the combinations of terms describing them.

Rules For the specification of rules we use property restrictions and associate them by owl:equivalentClass with the category (represented by owl:Class) for which the rule should apply. The SO includes three types of rules which are based on the mentioned three relations: 1. the rules for assembling composite terms from simple terms are based on the relation has_part and are associated with the subcategories of Composite_Search_Term (e.g., has_part_1 some Hemostasis_Adjective and has_part_2 some Hemostasis_Noun); 2. the rules for determining the maximal distance between simple terms within the composite term are based on the relation max_distance and are associated with the subcategories of Composite_Search_Term (e.g., max_distance value 2); 3. the rules for specification of term combinations describing a domain concept are based on the relation described_by and are associated with the subcategories of Search_Concept (e.g., described_by some Rebleeding_Phrase or described_by some Rebleeding_Noun).

2.2 Queries and OntoQueryBuilder (OQB)

We developed the OntoQueryBuilder, a software component which automatically generates Lucene queries, using the information contained in the SO. In the present stage of the SO and OQB, following parts of the Lucene query syntax are supported: single terms, simple term groups, proximity term groups, Boolean operators and brackets. We sketch the working method of the OQB. The queries are generated for all domain concepts (subclasses of Search_Concept, e.g., Rebleeding) which contain a property restriction, based on the property described_by (see Fig. 2). The Boolean operators are taken from the property restriction. Then we distinguish two cases. If one of the categories contained in the property restriction is a subcategory of Simple_Search_Term (e.g., Rebleeding_Noun) the labels of the instances must be simply connected by OR (rebleeding OR re-bleeding). If the category is a subcategory of Composite_Search_Term (e.g., Rebleeding_Phrase) in the first step the property restriction of the category, based on has_part, must be interpreted and contained subcategories of Simple_Search_Term must be identified. From the labels of the instances of these categories (Hemostasis_Adjective and Hemostasis_Noun) all combinations (possibly by using the order) must be assembled and connected by OR ("inadequate hemostasis" OR "incomplete hemostasis" OR "insufficient hemostasis" OR "inadequate haemostasis" OR . . .). In case of two categories, the instances of each category have altogether, e.g., 10 labels, 100 combinations must be assembled. These combinations can be automatically generated, which shows the benefit of this solution. If the maximal distance between the single terms is defined, the query will be adapted, accordingly ("inadequate hemostasis"~2).

3 Future Work

When compared to the traditional approach of using the ontology at index time [WD10], using the ontology at search time imposes some inflexibility on the information extraction. Even though this loss is more than compensated by the flexibility gain of not having to reindex on every ontology change, it makes sense to explore ways of transferring back parts of the information extraction to index time. Currently SO and OQB support basic full text queries with Boolean and NEAR operators. We plan to use information extraction technology (entity extraction, dependency parsing) to extract SO-independent entities and general relations from text. The results will be used to produce an enriched full text index, which in turn will be the basis for more precise query operations. We already implemented such an approach for named entities such as person names, organizations and locations and plan to extend this idea to storing general dependency relations.

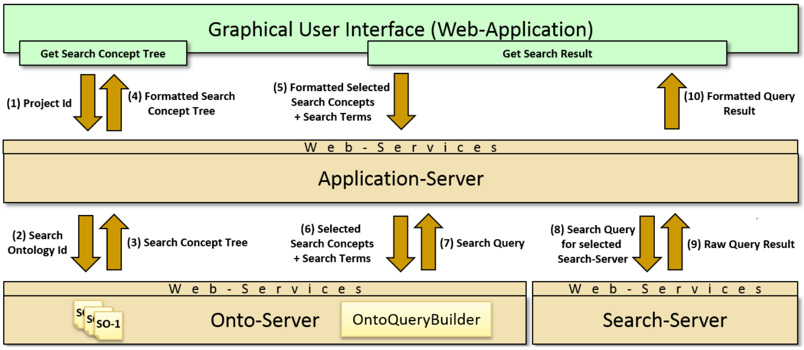
Graphical User Interface (Web-Application)
Get Search Concept Tree
Get Search Result
(1) Project Id
(4) Formatted Search Concept Tree
(5) Formatted Selected Search Concepts + Search Terms
(10) Formatted Query Result
Web-Services
Application-Server
(2) Search Ontology Id
(3) Search Concept Tree
(6) Selected Search Concepts + Search Terms
(7) Search Query
(8) Search Query for selected Search-Server
(9) Raw Query Result
Web-Services
SO-1
Onto-Server
OntoQueryBuilder
Web-Services
Search-Server

Figure 3: Prototypical Application

Moreover we are working on the prototypical implementation of the complete system as web application (see Fig. 3).

4 Related Work

There are various approaches which use ontologies or related/similar conceptions in full text search. [WD10] presents a survey of current ontology based information extraction systems, which all have in common the fact, that the ontology is used only in the text-processing step, and that the query answering system is decoupled from and unaware of the ontology. GoPubMed ([DDKS04]) is closely related to our work because it employs ontologies to search over medical domains, i.e., it also retrieves documents from PubMed. In contrast to our approach, GoPubMed is based on the idea that the ontology reflects the given documents and thus force the use of given taxonomies inherent to the system. Therefore, a user is not able to gain a higher precision determined by his subjective needs. Our approach solves this problem with the use of Search Ontology. Its inherent methodology simplifies the definition of search rules by domain experts. These rules enable the user to send powerful and complex queries which only have to be implicitly formulated.

Our approach of storing semantic information in a full text index is related to the approach used by Sindice[8]. However Sindice stores RDF triples in the index, while we use information extraction technology to identify potential entities and relations in documents and store the results in a full text index. No information from the ontology is used during indexing. In this way we do not have to change the index when the ontology changes.

5 Conclusion

The introduced Search Ontology provides a promising way for specifying and re-using complex search queries. It is mainly intended for domain experts such as patent examiners or retrieval experts in the field of post-market surveillance of medical devices. Advantages of this approach are: 1. the modular design of the Search Ontology which can be easily adapted and extended as well as its relatively simple structure, so that the Search Ontology is applicable by non-ontologist domain experts without excessive efforts and deeper knowledge of the query syntax, and 2. the fact that this approach accommodates constantly evolving ontologies, without the need to re-index the document collection on ontology changes. The rationale of the introduced ontology promises a feasible application and a sufficient flexibility to be used in a real-world application. The realization of the OntoVigilance project revealed that the domain experts may use the SO and are able to develop the corresponding domain ontologies. The first prototype of the OQB was tested and proved to generate correct Lucene queries.

6 Acknowledgment

This work was supported by the BMBF-project OntoVigilance, FK: 01-S12038D.

References

[DDKS04] Ralph Delfs, Andreas Doms, Er Kozlenkov, and Michael Schroeder. GoPubMed: ontology-based literature search applied to GeneOntology and PubMed. In *In Proceedings of German Bioinformatics Conference. LNBI*, pages 169–178. Springer, 2004.

[Her10] Heinrich Herre. General Formal Ontology (GFO) : A Foundational Ontology for Conceptual Modelling. In Roberto Poli and Leo Obrst, editors, *Theory and Applications of Ontology*, volume 2. Springer, Berlin, 2010.

[WD10] Daya C. Wimalasuriya and Dejing Dou. Ontology-based Information Extraction: An Introduction and a Survey of Current Approaches. *J. Inf. Sci.*, 36(3):306–323, June 2010.

[8] http://sindice.com/