Semantic Search in Heterogeneous Digital Repositories: Case Studies

Maria M. Nisheva[1,2], Pavel I. Pavlov[1], Peter L. Stanchev[2,3]

[1] Faculty of Mathematics and Informatics, Sofia University, Bulgaria
[2] Institute of Mathematics and Informatics, Bulgarian Academy of Sciences, Sofia, Bulgaria
[3] Kettering University, Flint, USA
marian@fmi.uni-sofia.bg, pavlovp@fmi.uni-sofia.bg,
pstanche@kettering.edu

Abstract. Semantic search augments the goals and scope of traditional search and information retrieval tasks. The paper analyzes the main aspects of semantic search and the corresponding features of some popular semantic search systems. The presentation is focused on a discussion of the implementation of a successful project in building modern semantic search engine. A brief analysis of the gained experience in some experiments with free software tools for advanced semantic technologies accomplished within the Master's degree course in Semantic Technologies at the Faculty of Mathematics and Informatics, Sofia University, is presented and the main techniques used for image and video semantic extraction are discussed in addition.

Keywords: digital repository, metadata, ontology, semantic annotation, semantic search, semantic interoperability

1 Introduction

During the last two decades, a large number of semantic search systems and various frameworks supporting the implementation of this type of systems were successfully built and put into operation.

Semantic search is aimed at improving the accuracy (precision and recall) of traditional search by understanding the user's intent and the right meaning of the particular terms and phrases that appear in the searchable repository or Web space. Semantic search systems use various sources and techniques to provide relevant search results: dictionaries of synonyms, thesauri, ontologies, context of search, intent, location, generalized and specialized queries, natural language queries, etc. Lately, modern semantic technologies enable the rapid development of new software for semantic search and information retrieval.

The paper discusses the main features of semantic search and presents the design and implementation solutions realized within a project directed to building tools for semantic search in a heterogeneous digital library. An analysis of the results of a series of experiments with free software tools for advanced semantic technologies accomplished in the form of course projects within the Master's degree course in Se-

mantic Technologies at the Faculty of Mathematics and Informatics, Sofia University, is presented in brief. The last section of the paper is aimed at a discussion of the most popular techniques, successfully used for image and video semantic extraction.

2 Main Characteristics of Semantic Search

Semantic search extends the scope of traditional search and improves its results using an understanding of the meaning of the search term(s) given by the user. Semantic search augments the goals of traditional information retrieval, which is mostly oriented to document retrieval, with additional entity and knowledge retrieval tasks [Guha, 2003]. It improves the conventional search and information retrieval methods by looking at the meaning of words that can be formalized and then described using ontology languages like RDF(S) and OWL [Wei, 2008]. In this way, a semantic search system is able to retrieve adequate results by reasoning on the user query and the accessible knowledge bases.

Five main directions of research and development in semantic search have been outlined in [Mäkelä, 2005]: augmenting traditional keyword search with semantic techniques, basic concept location, complex constraint queries, problem solving, connecting path discovery.

Augmenting traditional keyword-based search with semantic techniques is based on the use of proper domain ontologies or thesauri as sources for the user query expansion. In particular, the large WordNet ontology is often used for this purpose. First, the concepts referred by the given keywords are located in the ontology; then, a partial traversal of the graph representing the ontology structure is accomplished in order to find the terms, semantically related to the discovered concepts. These terms are utilized to either broaden or constrain the search.

Most semantic search systems are based on adding semantic annotations to data in order to improve search precision and recall on that data [Damjanovic, 2011; Mäkelä, 2012]. These semantic annotations are in fact metadata in the form of links to ontology concepts or individual descriptions in an appropriate semantic repository. The data the user is really interested in are individual objects that may belong to particular classes, but the domain knowledge is described primarily in terms of classes, their properties and relationships in the ontology. So if the properties point to objects, the search engine may ask the user to choose from the subsumption tree of classes in the ontology the class of instances (individual objects) he/she is looking for and then apply the constructed keyword-based filters to the properties of the corresponding set of instances.

Many semantic search systems provide tools allowing the user to formulate various kinds of complex queries as locating a group of individual objects of certain types connected by certain relationships.

Usually properties and property relations are used to traverse from a resource of interest to another but sometimes the discovery of paths in the graph connecting objects is the really interesting search result. Another use case associated with the vision of semantic search concerns describing a problem and searching for its solution by reasoning on ontological knowledge [Rui, 2008].

Thus the use of appropriate metadata and heterogeneous semantic knowledge (thesauri, ontologies, semantic annotations) is the main characteristic of semantic search in its various forms and scenarios. The application of ontologies is recognized as an instrument for explication of implicit knowledge [Cao, 2004] and as an approach to overcome the semantic heterogeneity of the searched repositories and datasets. Lately, some additional types of ontology applications directed for example to automatic reformulation of search queries to digital repositories containing semi-structured documents with incomplete and imprecise semantic annotations, ranking the reformulated queries, etc. [Mrabet, 2010] give an idea of the new trends in semantic search.

3 Implementation of Semantic Search Engine in a Digital Library with Bulgarian Folk Songs

The digital library DjDL preserves a collection of over 1000 folk songs from the Thrace region of Bulgaria. This collection is a part of the digitized archive of Prof. Todor Dzhidzhev presented in [Dzhidzhev, 2013]. The prototype of DjDL [Nisheva-Pavlova, 2011] was developed in 2010-2011 with the support of the Bulgarian National Science Fund. A new version of DjDL that has some substantial features of a social semantic digital library was designed and implemented in 2014-2015 [Nisheva-Pavlova, 2015].

The catalogue of DjDL consists of short descriptions of the particular folk songs included in the repository. These descriptions contain various types of metadata, for example: the title of the song, the song genre in accordance with different classification schemes, the region of folk dialect, the folklorist who gathered the song, the singer(s), the date and place of record, etc.

The folk songs treasured in the repository of DjDL are presented with their notes (musical notations), text (lyrics) and music (digitized versions of their authentic performances). DjDL has as its essential component a search engine that may be considered as a good example of semantic search engine.

The search engine of DjDL realizes two main types of search in the catalogue metadata and the lyrics of songs: keywords-based and semantic search. The semantic search tool provides a set of facilities for automatic reformulation (augmentation and refinement) of the queries for keywords-based search according to the available domain knowledge.

Two forms of conceptual domain knowledge are maintained in DjDL – a subject ontology and a set of concept search patterns based on this ontology.

The subject ontology contains definitions of a number of concepts, descriptions of their properties and several types of relationships between them, as well as a set of their representative instances. It describes an amount of knowledge in several domains, relevant to the content of Bulgarian folk songs: manner of life and family (professions, instruments, clothing, ties of relationship, feasts, traditions and rites, etc.), historic events, social phenomena and relationships, impressive natural phenomena. In addition, it contains a subontology of folk songs including various genre classifications of folk songs as well as a subontology of administrative division that combines the current administrative division of Bulgaria with the one from the beginning of the 20th century.

The so-called concept search patterns are natural language-dependent patterns of typical stylistic or thematic constructs frequently appearing in the lyrics of Bulgarian folk art. They are defined and have been used as domain knowledge aimed at providing satisfactory precision and recall of the search engine.

The semantic search tool of DjDL uses the subject ontology and the available concept search patterns in order to augment the search queries so much as possible, including in addition all discovered words and phrases that are semantically related to the one(s) originally given by the user. The resulting disjunction of concepts, their derivatives, synonyms and instances is properly visualized and placed at the user's disposal for optional refinement.

During the execution of a semantic search query, first of all it is augmented and refined with the assistance of the user. Then the lyrics of songs are consecutively searched and all documents with texts of songs containing phrases that are juxtaposed with at least one element of the augmented query are extracted. A list with the titles of the discovered songs is properly visualized on the user screen. When the user clicks on the name of a chosen song satisfying the augmented query, the text of this song is displayed in a new window along with the corresponding metadata. The discovered words and phrases that match the query are highlighted.

The search engine realizes some additional functionalities enabling the user to combine the search and retrieval of documents kept in the repository of DjDL with a kind of sentiment analysis of their texts. The sentiment analysis tool uses the subject ontology as a source of knowledge about the emotional intensity of its concepts and computes rough estimates of the mood of songs. For this purpose some ontology concepts are associated with appropriate positive or negative integers considered as sentiment estimates of the corresponding concepts. The sentiment estimates of the ontology concepts are used as default values for their specializations, forms and instances. The sentiment of a song is currently defined in accordance with the sum of the sentiment estimates of the particular words in the lyrics of this song.

The discussed search engine is developed as a client-server application built on the .NET Framework 4.5 and ASP.NET MVC 5. The tool used for its implementation is Microsoft Visual Studio Ultimate 2012 with additional packages for ASP.NET MVC 5. The SignalR library is used in the project in order to add real-time functionality to the software application. For JavaScript processing the jQuery library v. 1.10.2 has been used.

Figure 1 shows the software architecture of the digital library system of DjDL. The software implementation is based on Entity Framework 5 technology in combination with Code First. The current version of DjDL uses a local database (SqlLocalDB v. 11.0).

The class library RDFXMLClassLibrary was especially built for the purpose of automatic conversion of the original files with metadata and texts of songs (available in LaTeX format) to the RDF format in which they are processed by the search engine. LilyPond should be installed as an external software package on the server in order to generate files with standard musical notations and MIDI files with melodies of songs from the original source files treasured in the repository of DjDL [Kirov, 2012].

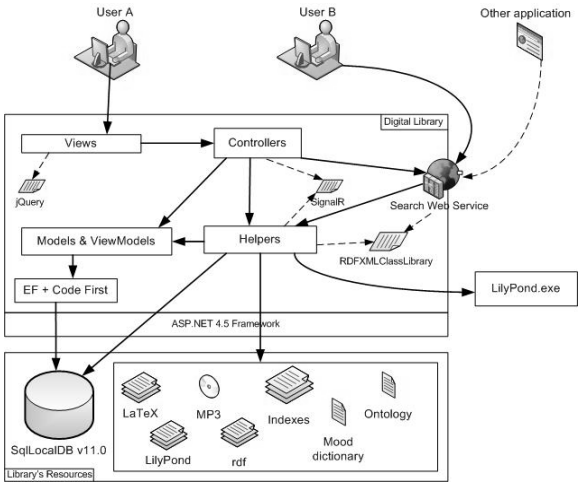

Fig. 1. Software architecture of the digital library system of DjDL [Nisheva-Pavlova, 2015]

The subject ontology is created using Protégé 4.3. Most concepts of this ontology are constructed as defined OWL 2 classes, by means of necessary and sufficient conditions formulated in terms of appropriate restrictions on certain properties. The current version of DjDL includes one more ontology (the Mood dictionary on Figure 1) containing some of the subject ontology classes and a "mood_estimation" property. The values of this property play the role of sentiment estimates of the respective concepts.

4 Experiments with Free Software for Advanced Semantic Technologies

Within the Master's degree course in Semantic Technologies at the Faculty of Mathematics and Informatics, Sofia University, a series of practical experiments with various free software tools for modern semantic technologies were performed in 2016 in order to obtain some firsthand experience of their functionalities and to design and implement small projects aimed at development of semantic databases in chosen domains, formulation and execution of different types of queries for search and reasoning on these databases.

The most promising results were obtained with the free version of GraphDB (http://graphdb.ontotext.com/). GraphDB may be characterized as a semantic repository – a graph database system that supports the load, preservation, management, and querying digital content in the form of semantically enriched datasets in real time. It uses ontologies to perform automatic reasoning about data and to create new facts that are implied in data.

We made more than 50 experiments with the restricted free version of GraphDB and realized that it gives the developer a good set of facilities for rapid implementation of semantic databases and semantic search tools adequate for the requirements and expectations of 92% of involved users. Within these experiments the size of the created and maintained semantic databases significantly changed. They were devel-

oped on the base of different sources – from relatively small RDF or OWL ontologies built for the occasion with the latest versions of Protégé to the DBpedia Ontology 2015 (http://wiki.dbpedia.org/services-resources/ontology) supplemented with specific DBpedia datasets.

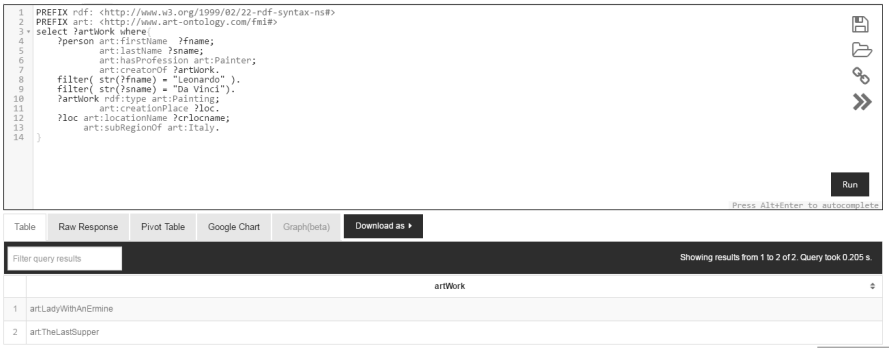

Fig. 2. An example query for semantic search in GraphDB Free [Todorova, 2016]

The only problem that should be overcome is the lack of interface module supporting the flexible and convenient (intuitive) construction of user queries for semantic search instead of the direct formulation of appropriate SPARQL queries (see for example Figure 2).

5 Semantic Multimedia Retrieval

Over the past decade, considerable progress has been made to make computers learn to understand, index and annotate multimedia, representing a wide range of concepts. The field of multimedia retrieval overcomes a major challenge: it needs to accommodate the obvious difference between the human vision and hearing systems, which has evolved genetically over millenniums, and the digital technologies, which are limited within pixels and waive capture and analysis. Typically, a content-based multimedia retrieval system consists of three components: feature design, indexing, and retrieval. The ***feature design*** component extracts the visual and waives feature information. The ***indexing*** component organizes the feature information to speed up the query or processing. The ***retrieval engine*** processes the user query and provides a user interface. During this process, the central issue is to define a proper feature representation and similarity metrics.

The main techniques used for image and video semantic extraction are presented in brief in the rest of this section.

MPEG-7 is an ISO/IEC standard developed by Moving Picture Experts Group. (http://www.chiariglione.org/mpeg). The standard describes the multimedia content data that supports some degree of interpretation of the information meaning. The visual descriptors in the MPEG-7 standard describe different aspects of the image content such as dominant colors, edginess, texture, etc. The MPEG descriptors are often

used in the process of image-to-image matching, searching for similarities, sketch queries, etc.

Local features are visual descriptors computed for local interest regions. Their typical applications include finding locations and particular objects, detecting image near duplicates and deformed copies. A drawback of the use of local features is that a single image is represented by a large set of descriptors. The most used local feature is the David Lowe **SIFT descriptor** based on four steps: Scale-space extrema, Key point localization, Orientation assignment, and Key point descriptor [Lowe, 1999].

Gabor filter (http://www.mathworks.com/help/images/ref/imgaborfilt.html ?requestedDomain=www.mathworks.com) is a very useful linear filter for edge detection. Frequency and orientation representations of Gabor filters are similar to those of the human visual system, and they have been found to be particularly appropriate for texture segmentation, target detection, fractal dimension management, document analysis, edge detection, retina identification, image coding and image representation.

Fisher Vector [Jaakkola, 1998] describes how the set of descriptors deviates from an average distribution, modeled by a parametric generative model. Fisher vectors have been applied in the context of image classification and large scale image search.

In the **VLAD** representation [Jégou, 2010] each local descriptor is associated to its nearest visual word in the codebook. The VLAD is the concatenation of the accumulated local features vectors.

The goal of **Bag of Words (BoW)** [Sivic, 2003] technique is to substitute each description of the region around an interesting point of the image with visual words obtained from a predefined vocabulary in order to apply traditional text retrieval techniques.

For aural features a commonly used descriptor is **power spectrum** – Mel Frequency Cepstral Coefficient [Muda, 2010]. For the motion features a commonly used descriptor is **dense trajectories** [Choi, 2014].

6 Conclusion

Our experience in the implementation of the discussed projects demonstrates once again that semantic search is a key issue in integration of heterogeneous databases and provision of semantic interoperability between heterogeneous digital repositories.

Various free software environments providing core infrastructure for modern semantic technologies can be used for the rapid development of flexible semantic search tools. Two important components of their user interfaces need some improvement at the present time: the ones for convenient query construction and presentation of results.

Acknowledgements. The presented work has been partially funded by the Sofia University SRF within the "Methods for Automated Semantic Annotation of Heterogeneous Datasets" Project, Contract No. 15/2016.

References

Cao S. et al. (2004). Semantic Search among Heterogeneous Biological Databases Based on Gene Ontology. Acta Biochimica et Biophysica Sinica 36(5), 365–370.

Choi J. et al. (2014). The Placing Task: A Large-Scale Geo-Estimation Challenge for Social-Media Videos and Images. In: Proceedings of the ACM Workshop on Geotagging and Its Applications in Multimedia (Orlando, Florida, November 2014).

Damjanovic V. et al. (2011). Semantic Enhancement: The Key to Massive and Heterogeneous Data Pools. In: Proceedings of the 20th International IEEE Electrotechnical and Computer Science Conference (Portoroz, Slovenia, 2011).

Dzhidzhev T. (2013). Folk Songs from Thrace. L. Peycheva, G. Grigorov, N. Kirov (Eds.), Sofia, Prof. Marin Drinov Academic Publishing House.

Guha R., McCool R., Miller E. (2003). Semantic Search. In: Proceedings of the 12th International World Wide Web Conference (Budapest, Hungary, 2003), 700-709.

Jaakkola T., Haussler D. (1998). Exploiting Generative Models in Discriminative Classifiers. Advances in Neural Information Processing Systems 11, MIT Press, 487–493.

Jégou H., Douze M., Schmid C. (2010). Improving Bag-of-Features for Large Scale Image Search. International Journal of Computer Vision 87 (May 2010), 316–336.

Lowe D. (1999). Object Recognition from Local Scale-Invariant Features. In: Proceedings of the Seventh IEEE International Conference on Computer Vision, Vol. 2, 1150-1157.

Kirov N. (2011). Digitization of Bulgarian folk songs with music, notes and text, Review of the National Center for Digitization 18, 35–41.

Mäkelä E. (2005). Survey of Semantic Search Research. In: Proceedings of the Seminar on Knowledge Management on the Semantic Web, Department of Computer Science, University of Helsinki.

Mäkelä E., Hyvönen E., Ruotsalo. T. (2012). How to Deal with Massively Heterogeneous Cultural Heritage Data – Lessons Learned in CultureSampo. Semantic Web 3, 85–109.

Mrabet Y., Bennacer N., Pernelle N., Thiam M. (2010). Supporting Semantic Search on Heterogeneous Semi-structured Documents. In: B. Pernici (Ed.), CAiSE 2010. LNCS 6051, 224-229.

Muda L., Begam M., Elamvazuthi I. (2010). Voice Recognition Algorithms using Mel Frequency Cepstral Coefficient (MFCC) and Dynamic Time Warping (DTW) Techniques. Journal of Computing 2(3), 138-143.

Nisheva-Pavlova M., Pavlov P. (2011). Ontology-Based Search and Document Retrieval in a Digital Library with Folk Songs. Information Services and Use, ISSN 1875-8789, 31 (2011), 157-166.

Nisheva-Pavlova M., Shukerov D., Pavlov P. (2015). Design and Implementation of a Social Semantic Digital Library. Information Services and Use, ISSN 1875-8789, 35 (2015), 273-284.

Rui H., Zhongzhi S. (2008). A New Approach to Heterogeneous Semantic Search on the Web. Journal of Computer Research and Development 45(8), 1338-1345.

Sivic J., Zisserman A. (2003). Video Google: A Text Retrieval Approach to Object Matching in Videos. In: Proceedings of the Ninth IEEE International Conference on Computer Vision, Vol. 2, 1470–1477.

Todorova G. (2016). Ontology of Art. Course Project in Semantic Technologies, Sofia University, Faculty of Mathematics and Informatics.

Wei W., Barnaghi P., Bargiela A. (2008). Search with Meanings: An Overview of Semantic Search Systems. International Journal of Communications of SIWN 3, 76-82.