Semantic Content Management and Integration with JCR/CMIS Compliant Content Repositories[*]

Suat Gönül
Software Research and Development Company
Silikon Blok Kat:1 No: 14
METU, Ankara, Turkey
suat@srdc.com.tr

Ali Anıl Sınacı
Software Research and Development Company
Silikon Blok Kat:1 No: 14
METU, Ankara, Turkey
anil@srdc.com.tr

ABSTRACT

Existing content management systems (CMSes) usually do not offer flexible, customizable means to create semantic, domain specific indexing and search mechanisms. Therefore, they either do not provide any semantic retrieval, search, browsing functionalities at all on the managed content or the semantic search functionality provided is limited as it depends on the manual annotation of content by users. So, in this study we describe a semantic content management flow by extracting implicit knowledge from both the structure of the CMSes and actual content within them. The task of additional semantic knowledge gathering and providing semantic operations on the content is a challenging task which includes adoption of several latest advancements in information extraction (IE), information retrieval (IR) and Semantic Web areas. In this study, we propose a new approach which provides automatic annotation of content managed in CMSes with the information retrieved from the Linked Open Data (LOD) cloud and several semantic operations on the content in terms of storage and search perspectives. We use a simple RDF path language to create custom indexes and retrive semantic knowledge from the LOD cloud suitable for specific use cases. All additional knowledge is materialized along with the actual content of document in dedicated indexes. This semantix indexing infrastructure allows semantically meaningful search facilities on top of it. We realize our approach in the scope of Apache Stanbol project, which is a subproject developed in the scope of IKS project, by focusing on document storage and retrival parts of it. We evaluate our approach in healthcare domain with different domain ontologies (SNOMED/CT, ART, RXNORM) in addition to DBpedia as parts of LOD cloud which are used annotate documents and content obtained from different health portals.

Categories and Subject Descriptors

H.2.5 [**Database Management**]: Heterogeneous Databases—*Data translation*; H.3.1 [**Information Storage and Retrieval**]: Content Analysis and Indexing—*Indexing methods, Thesauruses*; H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval—*Retrieval models, Search process*

General Terms

Algorithms, Design

1. INTRODUCTION

Upon the excessive increase on the data dealt with in the daily usage, content managements systems (CMSes) have gained an importance in terms of the managing the content and providing features that easies the end user's life. On the other hand, The LOD cloud[0] have become so wide that considering the existence of such a huge collection of data, enhancement of documents in CMSes with semantically related knowledge and building semantic features on top of semantically enhanced content is apparently reasonable.

However, today's CMSes do not exploit the LOD cloud in a flexible and automated manner sufficiently. Some of the web based CMSes such as Semantic Media Wiki[1], Drupal[2] provide semantic annotation of content by linking documents to the LOD cloud. In these systems, however, users are expected to add each annotation manually. On the other hand, in enterprise CMSes documents are represented with nodes. Pre-defined properties are associated to nodes to be populated manually. Although, such properties may give some explicit information about the content, without analyzing the content itself it is not possible to provide sophisticated semantic features. Also manually annotation of documents is an error prune and time consuming task[2].

In this paper we address the semantic incompetence of existing CMSes. We propose a unified approach which brings several latest advancements in different areas of information extraction (IE), information retrieval (IR) and semantic web technologies all together. We provide a mechanism for representing the JCR[3] or CMIS[4] compliant CMSes in RDF format. The representation is done by directly communicating

[*] This work has been part-funded by the European Commission under grant agreement FP7-ICT-2007-3/ No.231527 (IKS - Interactive Knowledge Stack).

[1] `http://semantic-mediawiki.org/`

[2] `http://drupal.org/`

[3] `http://www.day.com/specs/jcr/2.0/index.html`

[4] `http://docs.oasis-open.org/cmis/CMIS/v1.0/os/cmis-spec-v1.0.html`

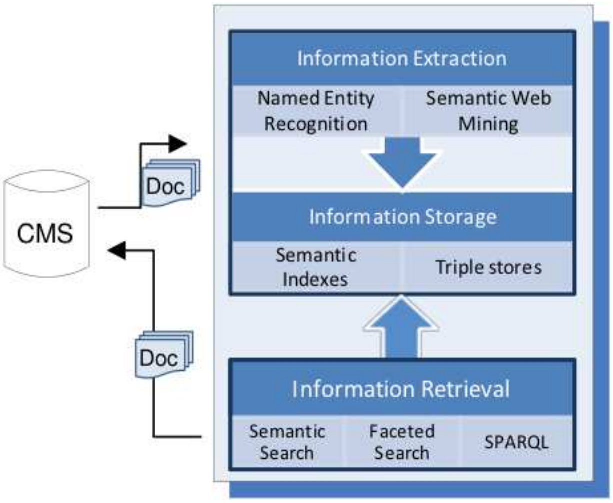

Figure 1: Semantic content management overview

with the underlying data model of CMSes and extract semantic relations, and generate RDF based semantic data to be processed in the semantic search process. The communication is bidirectional, thus conveniently structured RDF data can be pushed back to the content repositories.

Our approach also enables to create custom, semantically meaningful indexes tuned with specific domains, needs. This indexes are populated through the automatic annotation of documents with the knowledge to be obtained from the LOD cloud. By exploiting such a semantically enhanced indexing mechanism, we build a search machinery providing various ways to search for documents by keyword or structured queries or navigate on them using faceted search or related query terms.

The rest of this paper is organized as follows: We will give the overall definition of our approach in Section 2, then in Section 3, we elaborate the semantic indexing and search mechanism in detail. Then in Section 4, is about integration of CMSes with our approach. Related works will be given in Section 5. Before concluding, we will mention about evaluation of our approach in healthcare domain in Section 6 and finally the paper will be concluded in Section 7.

2. APPROACH

Not to interfere in the internals of CMSes, we propose an approach composed of integration of several techniques and tools from IE, semantic web mining, information storage and IR areas as external services to be used by CMSes.

Figure 1 summarizes the workflow between the CMSes and the semantic services provided by Stanbol. The flow starts with submission of documents from CMSes to the Stanbol framework. Using IE techniques, extra information about the content (e.g language, contained named entities etc...) are extracted. Later on, the extracted information is enhanced with the related semantic knowledge retrieved from the LOD cloud. All of the additional semantic knowledge related with the initial content is stored along with the initial content in customized, dedicated indexes, which are suitable for specific domain or use cases. Afterwards, several search functionalities in different modalities is provided over the indexed content and knowledge.

3. SEMANTIC INDEXING AND SEARCH

3.1 Semantic Indexing

To be able to submit documents to semantic framework, firstly the dedicated index should be ready. To create semantically meaningful indexes for specific needs, LDPath[5] language is used. LDPath is an RDF query language which is a valuable side-product of Linked Media Framework (LMF)[3] project. LDPath is also used to query entities through the LOD cloud.

Along with the actual content, different semantically related knowledge is stored in our approach. Stanbol framework allows enhancing of documents. Enhancements are returned in RDF format and they include various information about the content of the document such as language, named entities. While submitting a document to an index, the same LDPath which was used to create the specific index is used to obtain additional semantic knowledge from the LOD cloud for each named entity. Besides, this information is fully compatible with the target index.

Furthermore, Stanbol component provides configuration of specific datasets from the LOD cloud to be used in the enhancement process. This is an optional step, but if it is done once at the beginning, named entities are detected using the domain specific datasets and details of the entities are obtained from them. This would provide even more qualified annotations. After the initial configuration, the annotation and indexing process is fully automatic. During the document submission process, it is also possible to provide optional metadata in the form of **<field:value>** pairs. Even if the passed fields are not defined in the target index, they are dynamically indexed.

The crucial point of the proposed semantic indexing mechanism is the materialization of all related knowledge e.g knowledge retrieved from the LOD cloud or manually provided metadata along with the initial content. Having this valuable knowledge allows us to provide advanced search functionalities on top of them, which are described in Section 3.2.

3.2 Semantic Search

Exploiting the indexing mechanism, we build a machinery providing various search modalities such full-text search, document retrieval via structured queries e.g. SPARQL, Solr Query[6], document filtering with faceted search, and document exploration with related keywords suggested for the original query term. These search modalities are realized with the following services:

3.2.1 SPARQL Search

In the scope of Stanbol, enhancements of all documents are stored in a single RDF graph allowing querying of documents.

[5] `http://code.google.com/p/ldpath/`
[6] `http://wiki.apache.org/solr/SolrQuerySyntax`

3.2.2 Solr Search

This services makes possible to query the underlying Solr indexes by either keyword search or with Solr specific queries. Search requests are executed not only considering actual content, but also the related knowledge indexed along with the content.

3.2.3 Related Keyword Search

This service provides a related keyword suggestion mechanism for the initial query terms. Currently, related keywords are obtained three types sources which are Wordnet, any custom RDF data and dataset from the LOD cloud.

4. INTEGRATION WITH CONTENT MANAGEMENT SYSTEMS

In terms of integration with CMSes Stanbol framework provides functionalities for two main functionality, namely: *Index Feed* and *Bidirectional Mapping*. This functionalities can be used by all CMSes that are compliant with JCR or CMIS specifications.

4.1 Index Feed

This feature aims to synchronize the documents within a CMS with their counterparts which are semantically managed in the Stanbol framework. The synchronization is realized with two simple operations which are *submit* and *delete*.

During the submission process, properties of the documents in the CMS are also collected together with the content itself. Before indexing the content and its properties, enhancements regarding the content is also obtained. Eventually, all information is indexed and maintained within Stanbol. Figure 2 shows the detailed interactions which occur during the semantic content management process among the several Stanbol components.

4.2 Bidirectional Mapping

From one direction, this feature enables CMSes to represent their content repository structure in RDF format. This helps building semantic services (e.g reasoning facilities on top of the existing CMSes) using their RDF representations. Moreover, that representation can be used in the *Related Keyword Search* process in Stanbol. Therefore, it would be possible to navigate on the documents, considering the document hierarchy in the content repository.

From the other direction, bidirectional mapping feature makes it possible to exploit the LOD cloud within the CMSes. Apart from the already available data on the web, any RDF data can be mapped to the content repository. By mapping external RDF data, CMS items can be updated or new ones can be created. Thanks to this feature, already existing RDF datasets from various domains in the LOD cloud can be exploited to provide qualified classification/categorization for documents.

5. RELATED WORK

In this section, we give related studies with the work we present in this paper. Considering the way of integration of several defacto and novel technologies, our study differentiates from existing approaches. The study done in [1] describes a complete reference architecture for CMSes with semantic capabilities. This is the overall architecture covering our approach. As an extension to this architecture, we introduce methodologies to interact with JCR/CMIS compliant CMSes define in the section 4.

Configurable semantic bridges to extract semantics of JCR/CMIS compliant repositories into an OWL model based ontology are described in the scope of the work done in [4]. This approach proposes navigation of documents based on the class hierarchy in the generated ontology. However, this is a one-way approach which only provides extracting semantic information from a CMS but not feed the CMS with semantically enriched content. In our study, we allow update of the CMSes by external RDF data. Furthermore, our approach proposes various document retrieval, search and browsing methods apart from the ontology based navigation.

LMF[3] is a framework offering storage and retrival functionalities for media content. LMF supports annotation of media content using the LOD cloud, storage of the annotations along with the documents and publishes the stored content and its metadata in an interlinked manner. Thanks to the Solr field definition language, LMF allows creation of dynamically adaptable indexes for specific use cases. We use this dynamic index machinery as a base to storage and search capabilities in Stanbol. On top of that we propose more diverse document navigation features by means of related keywords retrieved from various sources such as WordNet, DBPedia or any external ontology. Besides, we provide an integration mechanism for CMSes.

Some of the web based CMSes such as Drupal or Semantic Media Wiki offer ways to integrate the LOD cloud and also provide semantic search through SPARQL. However, these system expects manual annotation from users. In our approach, the annotations are done automatically and they are easily configurable. In addition, we provide a flexible semantic index creation mechanism and several search modalities. Indeed, the aim is to provide more advanced semantic features to such CMSes like Drupal or Semantic Media Wiki. A usage of Stanbol framework with Drupal can be found in the article at [7]

6. EVALUATION

As the CMS to be supplied with semantic features, we have used a JCR compliant one, CRX[8]. Also to get health related enhancements for documents we use three health related datasets from the LOD cloud. To create an index suitable for health domain, an LDPath instance containing properties of terms/entities defined in the datasets was used.

Using the *Index Feed* feature, we have submitted all of the documents from CRX to Stanbol framework. Before indexing, documents are enhanced. Thanks to the health related datasets, we were able to recognize named entities such as diseases, drugs, adverse reactions, etc. The obtained enhancements were stored and indexed in the semantic index which is compatible with the enhancements themselves.

[7] `http://semantic-cms.info/`

[8] `http://www.day.com/day/en/products/crx.html`

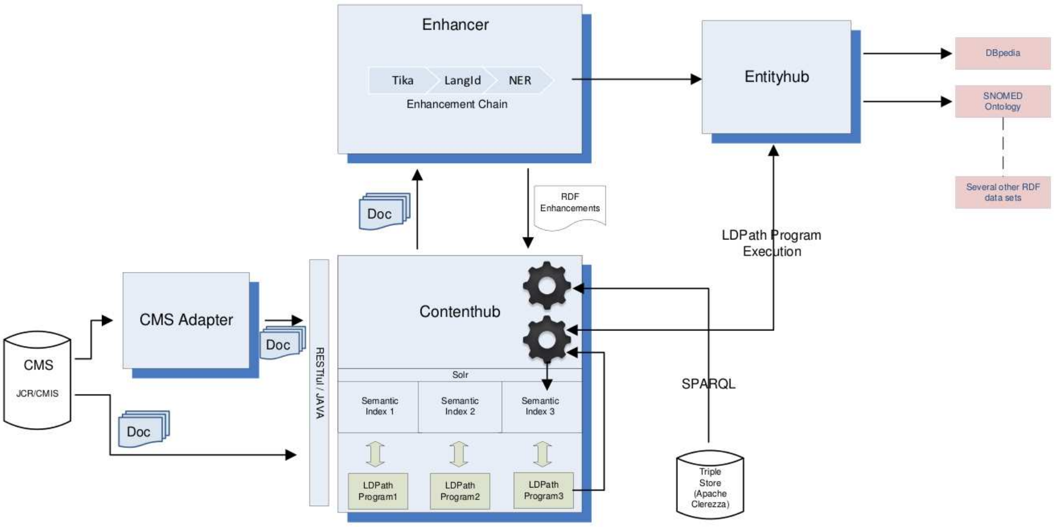

Figure 2: Semantic content management with Stanbol framework

We initially searched nonenhanced documents with several health related keywords. After that, we repeated the search with the same keywords on enhanced documents. In the latter case we obtained more related results. Because, some of search keywords were not included in the actual content of the documents but they were indexed as additional semantic metadata which is related with the documents.

Thanks to the domain specific enhancements and integration with health related datasets from the LOD cloud, we were able categorize documents according to health domain related facets. So, we were able to navigate on documents those facets. This categorization and navigation mechanisms did not existed in the actual CMS. Also, the enhancement process was fully automatic, once the external datasets were introduced to the system.

7. CONCLUSION

In this study, we have proposed a methodology offering semantic storage and retrieval services to be exploited by the CMSes which are not capable of managing documents together with their semantic information. The proposed approach has been realized in the scope of Apache Stanbol.

Stanbol uses Apache Solr as the underlying framework. Based on the Solr, it is possible to create semantic indexes which can be adapted for any specific domain through the LDPath language. Submitted documents are enhanced and their enhancements include detailed information about the named entities those contained in the document. Details of the named entities are retrieved by using the LDPath from the LOD cloud so that they would be compliant with the custom semantic index. All of the additional (semantic) knowledge is stored along with the document. On top of the indexes, in addition to full-text search structured queries through SPARQL or Solr Query syntax is also possible. Furthermore, documents can be navigated via the related keyword search mechanism.

In our approach, we propose a synchronization mechanism between the CMSes and Stanbol so that the CMSes can benefit from the functionalities of Stanbol. Also, Stanbol offers a mapping facility between the CMSes and external RDF data. Given any RDF data, this feature makes it possible to update the documents residing in the CMS. From the other direction, it provides the functionality of representing the structure of the CMS in RDF format so that actual structure of the CMS can be used in the semantic operations of Stanbol.

In this paper, we introduce a framework which makes use of the latest developments in the IE, IR and semantic web areas. The objective is to bring different, stand-alone implementations together and address the semantic requirements of the CMSes.

8. REFERENCES

[1] F. Christ and B. Nagel. A reference architecture for semantic content management systems. In *Proceeding of the Enterprise Modelling and Information Systems Architectures Workshop 2011 (EMISA'11)*, volume P-190, pages 135–148, Hamburg, Germany, September 2011.

[2] M. Erdmann, A. Maedche, H.-P. Schnurr, and S. Staab. From manual to semi-automatic semantic annotation: About ontology-based text annotation tools. In *Proceedings of the COLING 2000 Workshop on Semantic Annotation and Intelligent Content*, August 2000.

[3] T. Kurz, S. Schaffert, and T. Bürger. Lmf: A framework for linked media. In *Workshop on Multimedia on the Web (MMWeb) at the iSemantics Conf.*, pages 16 –20, Austria, sept. 2011.

[4] G. B. Laleci, G. Aluc, A. Dogac, A. Sinaci, O. Kilic, and F. Tuncer. A semantic backend for content management systems. *Know.-Based Syst.*, 23(8):832–843, Dec. 2010.