Knowledge acquisition, organization and maintenance for heterogeneous information resources

Nguyen G.
Institute of Informatics, SAS
Dubravska cesta 9
845 07 Bratislava
Slovakia
giang.ui@savba.sk

Laclavik M.
Institute of Informatics, SAS
Dubravska cesta 9
845 07 Bratislava
Slovakia
laclavik.ui@savba.sk

Babik M.
Institute of Informatics, SAS
Dubravska cesta 9
845 07 Bratislava
Slovakia
babik@saske.sk

ABSTRACT

The framework supporting data and knowledge acquisition, organization and maintenance for heterogeneous information resources is presented in this paper. It contains of the corporate memory (CM) and number of tools that work all together. The CM holds and manages documents, data and knowledge processed and created by tools. Tools work with data types e.g. documents, relational database and semantic data, etc. Each tool, in other view, can work also as independent tool to solve one specific problem. The framework finds its use in the automatic processing of documents with the aim to enable easy searching and finding information from wide space such as Internet. This whole chain process is quite complex and complicated with many non-trivial problems. The context of the knowledge management is based on the domain ontology which is a base for semantic data and creates a common background for entire system development.

1. INTRODUCTION

Nowadays, the Internet is becoming an universal repository of human knowledge which has allowed unprecedented sharing of ideas and information. Finding useful information is frequently a tedious and difficult. The difficulty is not only to know how to extract information, but also in knowing how to use it to decide relevance. The data retrieval process (as our project) aims to retrieving all objects which satisfy predefined conditions. At the moment, the approach of our framework is successful applied to the Job Offer search as the first pilot application of the NAZOU project [4], which enables users to find what they need, easily and adaptable to their profiles, preferences and requirements [2, 3]. A network of information is also provided and its services that can be processed by machines, which is different from the state of art, where web contain is mainly only human readable.

2. KNOWLEDGE CORPORATE MEMORY

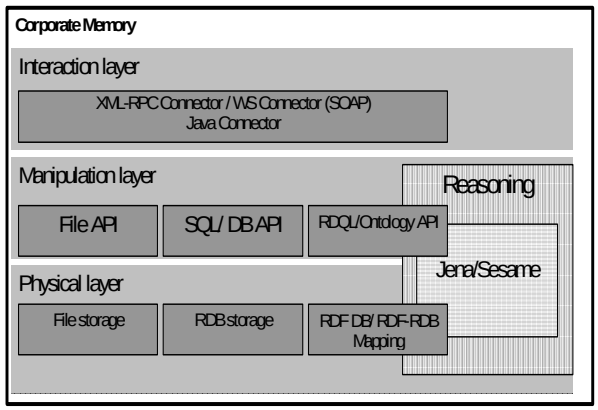

Figure 1: Cooperate Memory Architecture

Corporate Memory (CM) is accessible for other components using relevant client. The core of the CM is running as XML-RPC server and other components can call relevant client method via XML-RPC. CM is organized into three layers (Figure 1): physical layer (file system, database system, and ontological models), manipulation layer (access to the stored data and information) and interaction layer.

Ontology has become a very important aspect in many applications to provide a semantic framework to describe application domain. Ontology is a set of definitions of content-specific knowledge representation primitives (classes, relations, functions and constants). Ontology presents a shared understanding about a certain specific domain. There are also multiple inheritances, strong encapsulation, meta-data standards to train, discover and disambiguate meaning, and increased computing power. Although ontology enables processing knowledge and data, the most important role of ontology is in defining sharing meaning, emergence and discovery of gaps and for improving tacit knowledge transfer.

Semantic part of the CM is responsible for providing user interfaces for querying and manipulating the CM semantic content as well as providing the physical backend for persistent and transient storage of the semantics. The semantic model of the Web content is represented in the form of ontologies (OWL). The CM semantic part has two parts: the core interface (transparent access to the underlying knowledge repositories and reasoners) and the OntoClient inter-

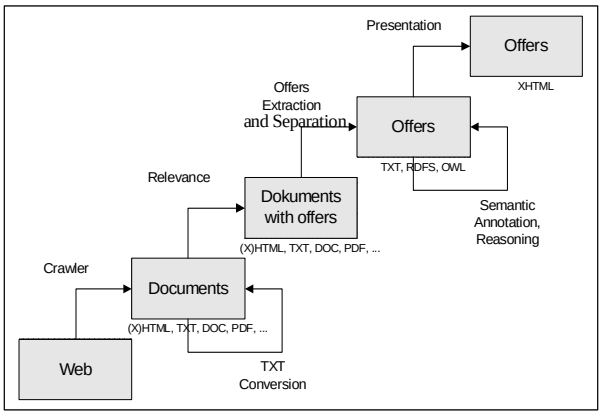

Figure 2: Chain of Tools

face (defines the possible interactions between the components of the system and the semantic part of the CM).

The relational DB management part of the CM is designed with virtualization concept, making actual DB system and DB connection object transparent to the client applications. The part of CM dedicated to file management provides a way for manipulating the file storage using unified application interface, making actual physical file storage transparent to the user or application. In the current implementation, CM's file storage is realized as a directory subtree of a file system directory tree. File management part of CM consist of core operations implementation and the client toolkit. Client toolkit can be configured to access the CM's file storage through local Java API with XML-RPC call or through Web Service interface, which is realized by OGSA-DAI framework.

3. DATA ACQUISITION, ORGANIZATION AND MAINTENANCE

Ontology based knowledge management includes activities like knowledge acquisition, creation, accumulation, sharing, reuse and capitalization. Knowledge items are abstracted to a characterization by metadata descriptions, which are used for further processing [1]. As it is described previously, here is a set of tools (Figure 2) that work with CM and each with other:

RIDAR (Relevant Internet Data Resource Identification): exploits the potential of existing search engines to identify relevant information resources on the Internet based on users-supplied search terms or more complicated search expressions. Details about identified resources (URL, title, etc.) are stored into databases.

WebCrawler traverses identified resources by RIDAR and downloads pages. These pages are then analyzed by ERID (Estimate Relevance of Internet Documents) tools, which estimate the page's relevance. The relevance estimation tires to decrease amount of downloaded documents by eliminating the pages with uninteresting content

DocConverter and DocIndexing transform documents from one to another format. At the moment, it transforms HTML to TXT documents for the need of other tools. The tool is accessible through WSRF standard compliant Web Service (WS) interface using OGSA-DAI framework. WS interface facilitates integration of the tool in distributed, heterogeneous environment.

ExPoS (Offer Extraction) processes downloaded and converted documents with offers and removes irrelevant information such as advertisements, etc. using several noise analysis methods. OSID (Offer Separation) separates blocks of offers from documents that contain more offers according to structure and offer identification indications. These two tools closely work together, they both deal with text processing and text analysis problem that have many non-trivial features. Their output is very important and userful for knowledge acquisition and organization.

Ontea (Ontology Based Text Annotation) annotates text version of offers by ontology individuals via regular expressions as relevant semantic properties of the offer then creates ontology form of offers according to predefined ontology. This can help e.g. in categorization, common visualization of documents, searching and knowledge inference or reasoning. While most of annotation solutions try to find and create an object in text or to provide semantic tags for a reader, Ontea tries to detect ontology elements within the existing domain ontology model. In this stage, experiments show the archived success of the tool around 80%.

4. CONCLUSIONS

In this paper, the chain of tools and functionalities of the CM of the framework is presented. Tools are almost independent but are integrated all together around the CM to achieve common aim as the whole. The approach is widely used in EU Research and Development and national projects for automatic data processing for knowledge management. At the moment, the work of our team is concentrated on enrichment and improvement of functionalities of tools and CM as well as the cooperation among them.

5. ACKNOWLEDGMENTS

This work is supported by NAZOU SPVV 1025/2004, EU RTD IST K-Wf Grid FP6-511385, APVT-51-024604 and VEGA No. 2/6103/6.

6. ADDITIONAL AUTHORS

Additional authors: Ciglan M. and Gatial E. and Balogh Z. and Oravec V. and Hluchy L.

7. REFERENCES

[1] R. Beaza-Yates and B. Ribeiro-Neto. *Modern Information Retrieval*. ACM Press, Addison Wesley Longman, 1999.

[2] M. Bielikova and J. Kuruc and V. Marko. Entry into virtual university space through web-based e-application. *Elfa*, pages 403–410, 2004.

[3] R. Lencses. Indexing for the information retrieval system supported with relational database. In *Sofsem?05 Communications*, 2005.

[4] P. NAZOU Team. Nazou project website. In *http://nazou.fiit.stuba.sk/*, 2006.