Acquiring, Organising and Presenting Information and Knowledge from the Web

Pavol Návrat, Mária Bieliková, Viera Rozinajová

Abstract: *This paper describes aims, progress and some results of a research conducted in a project that is aimed at devising ways of processing of information and knowledge in a heterogeneous environment, in particular at acquiring, organising and presenting information and knowledge from the web. Important part of the project are pilot applications. Their main purpose is to test the devised methods and tools as extensively and realistically as possible. From the two pilot applications that are planned, we have already developed the first one, devoted to offers and requests on the job market. This application allows to incorporate most of the tools. They show usefulness in supporting the user when looking for new job positions, when mediating services connected with employment or when selecting an applicant for a job for various job positions.*

.

Key words: *Semantic Web, Ontology, Supply, Demand.*

Introduction

The problem of acquiring and presenting information and knowledge from the web is a topic of intensive research interest nowadays. Many research groups at various places attempt to tackle it from different perspectives. Here we can mention for instance project AKTORS supported by the British government (www.aktors.org), projects supported by European Union, i.e. On-To-Knowledge (www.ontoknowledge.org), REWERSE (rewerse.net), Knowledge Web (knowledgeweb.semanticweb.org). Project SIMILE (simile.mit.edu) is a result of cooperation of a consortium built from W3C, MIT Libraries and MIT Computer Science and Artificial Intelligence Laboratory. There are also other projects on the topic, some of them going on in Slovakia.

When talking about research activities concerning this area in Slovakia, we must admit that they are rather fragmentized. It has been the project "Tools for knowledge acquisition, organization and maintenance in the environment of heterogeneous information sources" funded by governmental program of research and development "Building information society" that finally provided supportive conditions for forming a group of several dozens of researchers, mostly young ones, which started to cooperate very intensively. These researchers are from Slovak University of Technology, Faculty of Informatics and Information Technology - the main contractor of project, then from Institute of Informatics, Slovak Academy of Science, from University of Pavol Jozef Šafárik, Košice and from Softec, Ltd., a private enterprise in the IT sector. The project is significant from two points of view: (1) research level – the topic is very up-to-date, and (2) level of integration of research activities in Slovakia, facilitating cooperation of a relatively large group of researchers.

.

Research areas

Coming out from the main research objective of the project, which is improvement of providing actual and relevant information from web, the project is focused to the investigation of new ways of information and knowledge processing in the environment of heterogeneous sources, especially with imperfect and vague information. In the coincidence with the described goal, the research is oriented towards these areas:

- Models of heterogeneous environment (uncertainty, systems for modeling imperfect information, models of application domain, user models, context models, navigation models, metadata and ontologies, multilanguage approach, multiagent systems,
- Knowledge acquisition (information recommendation, obtaining of user or environment model, special languages for flexible query, ontology creation),

- Knowledge organization (ontologies, various inductive methods, clustering, indexing, small world networks),
- Knowledge presentation (adaptive navigation, adaptive content presentation).

Application domain

The project covers application domain of providing information and knowledge about job offers. Social importance of this application is undoubtedly high and there is no need to argue about it. It is a typical problem, and its nature is heterogeneous in more dimensions. The sources, which contain offer are different –ranging from global enterprises to small local companies. The way of presentation will also vary. In the European Union we have more than a dozen of official languages, and there are naturally many more languages that are used on web. Job offers in these languages can be just as useful as the ones written in official languages. Other dimensions of hetergoneity are for instance profession, region etc. Professions may have different customs, ways of expressing, idioms and patterns of describing what is offered and what is sought. Similarly, natural cultural differences among regions can influence the meaning of offers and requests.

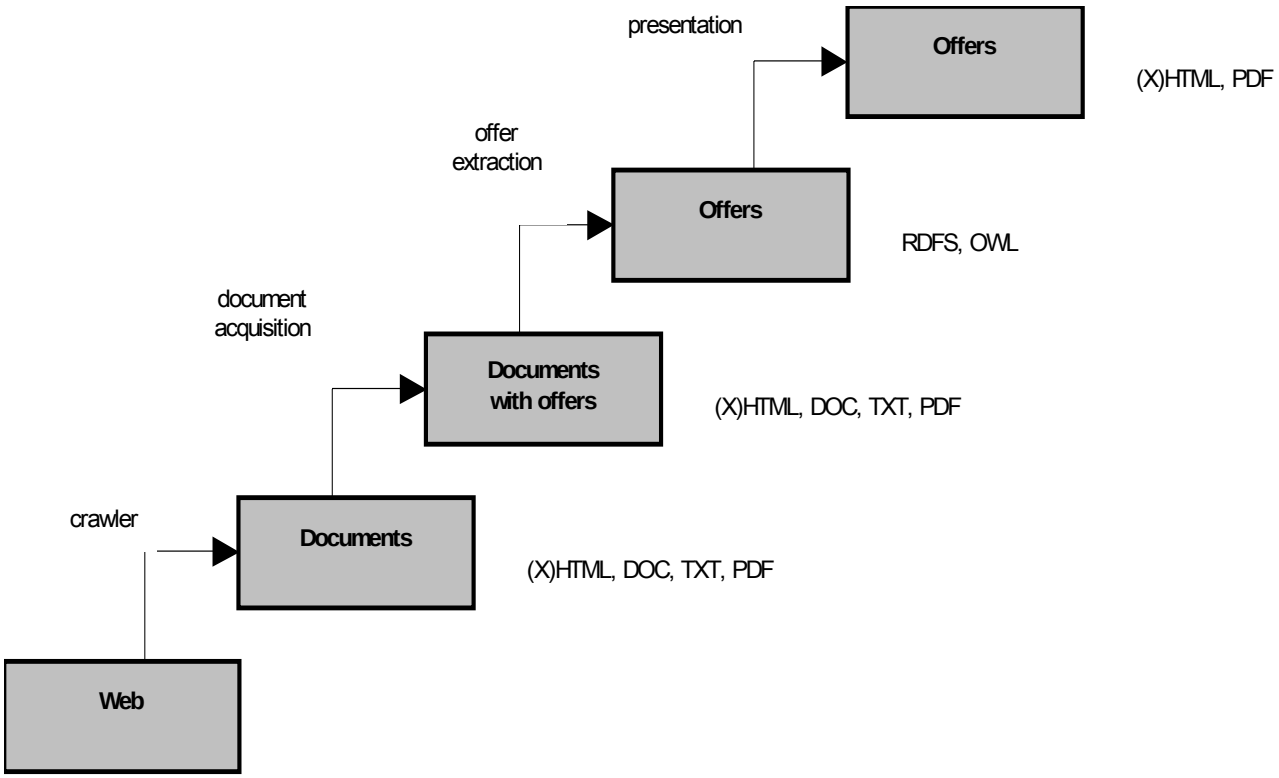
In the "jungle" of many data on the web, one has to search, make accessible and present to the user those informations, which best suit his/her need or preferences. Informations provided must be up to date, and this aspect should be watched by an agent working in the background. Finally the most important aspect of the whole project is the degree of improvement of the access to the information and knowledge about job market, which helps him find the most suitable job for him.

The main task of the pilot application is to search effectively information and knowledge about job offers and to mediate them to the potential applicants for these positions, and thus:

- Improving the process of selection of suitable job for the applicants by more effective processing of job offers,
- Increasing chances of the applicant to find the job, which best suits his requirements and possibilities by following and adapting the retrieved information, also by providing the sources which were hidden so far,
- Enabling employers to find appropriate job applicants which fulfill their requirements.

From the data on web to information and knowledge for the user

It is quite difficult to describe all aspects of the project in a few paragraphs. Some picture could be provided by describing the designed tools and transformation performed by them from the data acquired from web to information and knowledge presented to the user. The tools realise sequence of data acquisition and processing, so that they operate on various levels of semantics understanding of individual sources. The sequence follows in successive steps form acquiring data containing job offers from Internet, through identifying documents in which offers are incorporated, offers extraction, their organizing up to their presentation to the user. This could be characterized as transformation of the part of the web to the semantic web, where existing documents are transformed to such a representation, which augments at the highest level presented information by semantic concepts utilised with advanage by their automated processing.



Transformation data from web to presented information

Design of components for tools was influenced also by the fact that software systems in general and especially those concerned with hypermedia are developed without special attention to organized reuse, which is one of the main conditions of automating software development. Ad hoc reuse has not approved to be useful in practical applications. Development for reuse, which is the basis for domain-oriented approaches, is inevitable. (Smolarova 97, Vranic, 05). These approaches focus their attention to the whole domain of application, which must be appropriately constrained. Architecture and components for the whole family of software systems in the given domain create outcome of this development process, and the specific software systems are being developed in this way. This area is also the subject of research in our project.

Tools for the acquisition of data and offers

- the tool for the offers' resources identification on the Internet (RIDAR, Relevant Internet Data Resource Identification),
- the tool for downloading offers (WebCrawler),
- the tool for the estimation of relevancy (ERID, Estimate Relevance for Internet Documents),
- the tool for the extraction of individual offers (OSID, Offer Separation for Internet Documents),
- the tool for the conversion of documents (DocConverter),
- the tool for editing ontology (JOE, Job Offer Editor)

Tools for the analysis and the organization of data and offers

- the tool for the annotation of documents (OnTeA, Ontology-Based Text Annotation),
- the tool for fuzzy conceptual clustering (OSFCL, One-Sided Fuzzy Concept Lattice Clusterer),
- the tool for probabilistic clustering of documents (Aspect),
- the tool for navigation in the map of firms and job positions extracted from job offers (Job Cluster Navigator),
- the tool for classification by the usage of inductive logical programming (IGAP),
- Top-K aggregator (TopK),
- the tool for related words management (RWM, Related Words Manager),
- the tool for fulltext indexing of text documents and for retrieval (DaiRTFS, Data Access and Integration Rich FullText Search),
- the tool for fulltext indexing and search in documents (JDBSearch),
- the tool for search of offers with criteria (CriteriaSearch).

Tools for the presentation

- frame for the presentation of offers represented by ontology (Prescott, Presentation-Cocoon-Ontology)
- faceted browser (FACTIC, Faceted Semantic Browser)

Conclusions and future work

One of the main outputs of the project was an internet portal, which serves for verifying research results. This is valuable especially in experimenting with new ways of knowledge processing in environment of heterogeneous information sources. Problem domain is the area of job market. This Internet portal has been created by means of tools for knowledge acquisition, organization and maintenance. In this way we have practically approved methods and techniques designed within the research project The main advantages of our approach could be summarized as follows:

- use of ontology (reasoning, document similarity)
- active search of information about job opportunities (checking company portals)
- concentrating job offers from heterogeneous environment
- robustness from the point of view query formulation

Original design of software architecture for intelligent knowledge management as well as original approach to component development for tools that acquire, process and maintain information are also important achievements of this project. These tools are in various development phase, some are in stage of development, some in stage of specification, others in stage of design and some of them already in stage of verification. Their integration into software architecture is gradual. We develop paralelly the experimental ontological base of job offers assuming that this ontology will be usable in other similar oriented projects.

Acknowledgements

This work was partially supported by the Slovak State Programme of Research and Development "Establishing of Information Society" under the contract No. 1025/04 and the Scientific Grant Agency of Republic of Slovakia, grant No. VG1/3102/06.

REFERENCES

[1] Bieliková, M., Kuruc, J. Sharing User Models for Adaptive Hypermedia Applications. In H. Kwasnicka, M. Paprzycki (Eds.), Proc. of ISDA 2005, Sept. 2005, Wroclav, Poland, IEEE Computer Society Press, Loa Alamitos, pp. 506-511.

[2] Bieliková, M., Grlický, V., Kuruc, J. Rámec pre prezentáciu informácií reprezentovaných ontológiou. In P. Vojtáš (Ed.), Proc. of ITAT 2005 - Workshop on Theory and Practice of IT, Sept. 2005, Račkova dolina, pp.325-334.

[3] Filkorn, R., Návrat, P. Feature-based Filtering in Semantic Web. In B. Thalheim and G. Fiedler(Eds.), Emerging Database Research in East Europe, Proc. of the Pre-Conference Workshop of VLDB 2003, pp. 46-50.

[4] Frivolt, Gy., Bieliková, M. An Approach for Community Cutting. In V. Svátek, V. Snášel (Eds.), RAWS 2005 - Proc. of the 1st Int. Workshop on Representation and Analysis of Web Space, Sept. 2005, Prague, Czech Republic, pp. 49-54.

[5] Gatial, E., Balogh, Z., Laclavík, M., Ciglan, M., Hluchý, L. Focused web crawling mechanism based on page relevance. In P. Vojtáš (Ed.), Proc. of ITAT 2005 - Workshop on Theory and Practice of IT, Sept. 2005, Račkova dolina, pp.41-46.

[6] Gurský, P., Horváth, T. Dynamic search of relevant information. In L. Popelinsky (Ed.), Proc. Znalosti 2005, Vysoke Tatry, Slovakia, pp. 194-201.

[7] Horváth, T., Krajči, S., Lencses, R., Vojtáš, P. An ILP model for a graded classification problem. J. Kybernetika, 40 (2004) pp. 317-332.

[8] Laclavík, M., Gatial, E., Balogh, Z., Habala, O., Nguyen, G., Hluchý, L. Semantic annotation based on regular expressions. In P. Vojtáš (Ed.), Proc. of ITAT 2005 - Workshop on Theory and Practice of IT, Sept. 2005, Račkova dolina, pp.305-306.

[9] Polčicová, G., and Návrat, P. Semantic Similarity in Content-based Filtering. In Manolopoulos, Y. and Návrat, P. (Eds.), Proc. of ADBIS 2002 - Advances in Databases and Information Systems, Springer LNCS 2435, pp. 80-85, 2002.

[10] Polčicová, G., Tiňo, P. Making sense of sparse rating data in collaborative filtering via topographic organization of user preference patterns. In Neural Networks, Elsevier, Vol. 17, (2004), pp. 1183-1199

[11] Smolárová, M., Návrat, P. Software Reuse: Principles, Patterns, Prospects. Journal of Computing and Information Technology, 5(1997), 1, 33-48.

[12] Vranić, V. Multi-Paradigm Design with Feature Modeling.ComSIS (Computer Science and Information Systems), Vol. 2, No. 1, (2005), pp. 79-102.

ABOUT THE AUTHORS

Prof. Pavol Návrat, PhD., Prof. Mária Bieliková, PhD., Viera Rozinajová, PhD., navrat@fiit.stuba.sk, from:

Institute of Informatics and Software Engineering, Faculty of Informatics and Information Technologies, Slovak University of Technology in Bratislava, Slovakia.