

Intelligent Information Processing in Semantically Enriched Web

Pavol Návrat, Mária Bieliková, Daniela Chudá, and Viera Rozinajová

Institute of Informatics and Software Engineering
Faculty of Informatics and Information technologies
Slovak University of Technology, Ilkovičova 3, 842 16 Bratislava
{navrat,bielik,chuda,rozinajova}@fiit.stuba.sk

Abstract. Acquiring information from the Web is a demanding task and currently subject of a world-wide research. In this paper we focus on research of methods, and experience with development of software tools designed for retrieval, organization, presentation of information in heterogeneous data source spaces such as the Web. We see the Web as a unique evolving and unbounded information system. The presented concepts can be used also in other specific contexts of information systems in organizations that increasingly become worldwide and weaved together considering information processing.

1 Introduction

The World Wide Web allows any kind of information (of any content, but also of almost any media: text, picture, sound) possible to be “put on the Web” and be read by anyone anywhere in the world. All what we put on the Web can contain a reference to other information present on the Web. And the Web can interpret these references so that in case we want it, the referred information is accessed.

We structure the information we put on the Web into pages and pages into sites. That is of course just the technical level of structuring. By mutual references between pages regardless of whether they belong into one site or whether they belong to one author, there can be formed all sorts of connections, expressing relations between information listed on the particular pages. The described technical level of structuring gives presumptions for forming social networks of people communicating with each other through their personal computers. Such computers together with software that enables variety models of communication in some community (e.g., discussion groups) are becoming increasingly also social computers.

In this paper we rely upon our research and experience with the design of methods and development of software tools designated for acquisition, organization and presentation of information and knowledge from large data spaces employing the semantics either explicitly defined or discovered. This research was a part of the research project called NAZOU (“Tools for acquiring, organization and maintenance of knowledge in heterogeneous data sources space”),

nazou.fiit.stuba.sk) [16]. The main goal is improvement of providing current and relevant information from the Web by automatic processing.

Throughout the paper we feature examples of application domain of the mentioned project – the domain of acquisition, organization and presentation of job offers. That does not mean that the described approaches can be used exclusively for the domain of job offers. Most of the devised methods are applicable also in other domains mainly connected to information processing in organizations.

2 Related Work

The problem of information processing is the subject of intensive study worldwide [15]. Especially the idea of the Semantic Web inspired several research groups aiming at effective information processing on the Web. The Semantic Web as “a Web of actionable information – information derived from data through a semantic theory for interpreting symbols” [20] gives an opportunity to reason on documents and convert them automatically through data to information.

Since Tim Berners-Lee presented in 2001 a vision of the Semantic Web, several research projects based on this idea started. Here we can mention especially project AKT – Advanced Knowledge Technologies that has been financed by the British government (www.aktors.org), past and current projects supported by the European Union, e.g. REWERSE – Reasoning the Web (rewerse.net), KP-Lab – Knowledge Practices Laboratory (www.kp-lab.org), K-Space – Knowledge Space of Semantic Inference for Automatic Annotation of Multimedia Content (www.k-space.eu), On-To-Knowledge (www.ontoknowledge.org), Knowledge Web (knowledgeweb.semanticweb.org). SIMILE – Semantic Interoperability of Metadata and Information in unLike Environments (simile.mit.edu) – joint project conducted by the MIT Libraries and MIT CSAIL covers several projects aimed at developing open source tools that empower users to access, manage, visualize and reuse the Web content.

There are more projects dealing with semantically enriched data processing in large spaces. Typically, they use ontologies as a base for metadata representation and reasoning, mostly employing RDF/OWL W3C recommendations for representation [7] and deal with issues of ontology querying as a kind of information retrieval [12]. They define new ontologies either domain dependent [6, 19] or domain independent [11] and work on tools for ontology specification and maintenance [1]. Most of the projects consider a user as an important stakeholder and research or just employ techniques for personalization [5]. Most of the mentioned projects face the problem with non existing fixed data collections that would serve for experimental comparison of particular approaches to information processing in the Web (such as TREC, trec.nist.gov).

We present a concept that aims to cover the whole process of information processing in large data spaces. We provide methods for solving particular problems. Even though the methods cannot cover all aspects of the “information processing problem” they present contribution by providing a consistent chain of information processing that can be reused in several application domains.

3 From documents to information for the user

Coming up from the main goal, which is improvement of providing relevant information from the Web to a user in a way as much automated as possible, we focus on describing approaches in the research of new methods and tools of information processing in large data spaces. Research in this area naturally incorporates creation of models of heterogeneous environment. Data, which we have at disposal, are of uncertain nature and provide us with imperfect information. In connection with retrieving and processing of information that is relevant for the individual user in the given context we not only need data models of the content, but also user and context models.

Our approach is based on forming a procedure that involves software tools, which carry out methods for information processing, transform data acquired from the Web documents to information and knowledge, presented to the user. Software tools carry out the sequence of data acquisition and its processing to information, and thus they have to work on various levels of semantic understanding of individual data sources. For example, on the level of acquisition of data from the Web they work with partly structured text, for which they estimate its relevance with respect to the domain (e.g., whether it is a job offer) and estimate the value of individual parts of the text, or of the whole document (e.g., what is the company name, in what area the job offer is, as on the contemporary Web this information is by no means specified). When organizing the acquired offers, we use the already discovered document characteristics and we form for example a network of companies, which provide job offers in respective areas and estimating similar job offers based on that.

Information processing is accomplished in several steps, sequentially linked with each other, from selection (filtering out) and document acquisition from the Web, which comprise required data (in our case data on job offers), through identification and annotation of those documents that really comprise job offers (in order to extract required data); selection of individual job offers; their analyses and organization to their personalized presentation to the user (see Fig.1).

We transform a part of the Web to the Semantic Web. The existing documents are transformed into a representation, which describes their content using metadata. In such a way, the whole information becomes suitable for automatic processing. Naturally this process is iterative (even if its skeleton is formed as a sequence). This sequence is in fact provided many times repeatedly based on processing of atomic document or part of data. Considering individual tools that realize particular tasks they should synchronize each other to be able to process the required output for the user. The concept of the Web itself is advantageous here – data constitute a basis for tools integration.

4 Intelligent management of information in the domain of job offers

Suppose there is a user, who (with support of web services) is looking for information about something he is interested in. Considering our domain we suppose

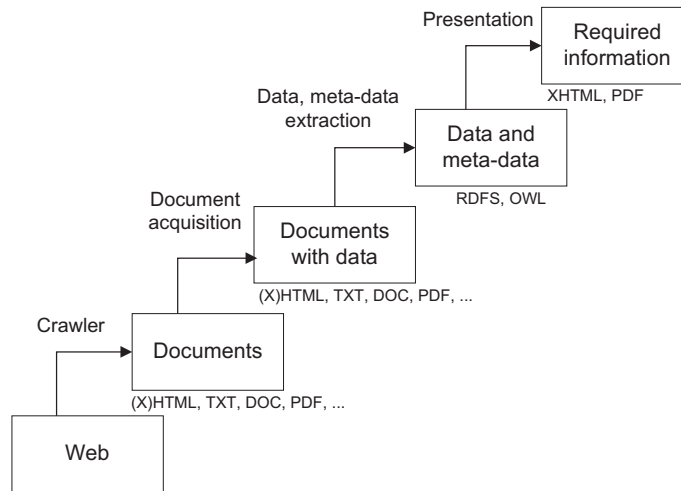


Fig. 1. Transformation of documents to information and knowledge to be presented.

there is a user, who is looking for a job with a good salary, reasonable distance from his permanent address and one that corresponds to his education and abilities (and possibly other expectations). For the common problem of retrieving information, the task was to design methods for processing data so that by means of the tools realizing these methods it was possible to construct an information system, which will help the user retrieve relevant information.

Immediately some questions arise like e.g., in what form will the user communicate with the system? What will be prepared before user querying (i.e., can tools do some preprocessing that would help increase effectiveness of information processing)¹? How to prepare for potential arbitrary requirements or expectations of the user? Are we going to offer to him rather a closed system and means enabling querying based on examples or are we going to devise a system more open to individual requirements? How to settle up with continuous changes of data space (job offers originate, some change, or simply become obsolete)?

Indispensable is the phenomenon of user preferences. The concept of a good salary can for various users be different. Reasonable distance from his permanent address is also a concept, which is differently perceived by a wage-earning mother and differently by a young person, who is becoming independent. It can also happen that two different job offers are incomparable. One is better in salary parameters and the second one can be better regarding distance. An answer should not only contain all (possibly also relevant, or interesting) offers for the particular user, but should be arranged according to their relevance.

¹ While designing methods for information processing we must have regard to the fact that we work with a data space that contains large number of information, more than we are able to process with contemporary means in real time. At the same time the content of the Web constantly changes, which requires a compromise between acquiring “some” relevant information.

That brings us to one possible view of the information system working in semantically enriched document space (see Fig. 2). We distinguish in it web sources (represented by documents, be it static or generated from the hidden Web), and components that acquire relevant documents and process them to be presented for human users. Every proposed representation of documents introduces for a large data space such as the Web a definite loss of information (as it is neither possible, nor effective or realistic to download and save entire web). However the matter is that we choose representation with a reasonable loss of the content. The tendency is even to look for mechanisms with a resulting negative loss, i.e. we enrich sources, which we acquire from the Semantic Web, (ideally) automatically amend to individual terms or parts of documents meaning by metadata.

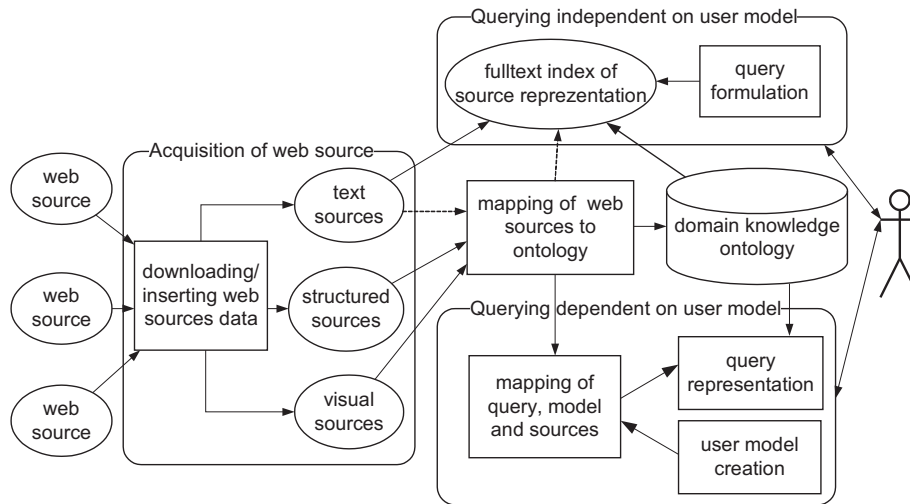


Fig. 2. Acquiring and using data from web sources [17].

Categorical knowledge of application domain – in our case job offers (represented in the Semantic Web applications often by ontologies) are necessary to provide querying on acquired data with the aim of support of search and navigation within the information space (data space of acquired from the Web is transformed to information space now).

The important component is “querying based on the user model” – a concept of personalized information presentation. Personalization can be achieved in such a manner that the recommendation depends only on the user activities (e.g., clicking while browsing) or so that it considers also additional information about the user represented in user model (such as user background) [2], or also on information about other users (employing social relations). Necessary presumption of methods for adaptation of the content and navigation is the possibility

of comparison of individual investigated and presented entities (in our case job offers or their parts in the sense of used representation).

We assume the following sequence of data processing with the goal to provide effectively relevant information (we feature on example of job offers, although the sequence is more general):

- primary documents on the Web,
- acquired documents that contain job offers,
- documents containing relevant data concerning the task of retrieving job offers,
- extracted job offers from the documents identified as relevant,
- job offers (or parts thereof) presented to a particular user or a group of users.

Implementation of the described sequence in view of the presented concept of intelligent management of information can be divided into three relatively independent tasks, which however are interlinked and mutually influence each other: (i) document and data acquisition, (ii) analysis and organization of data and information, (iii) personalized presentation of information including methods for creating and maintaining the user model.

In the following subsections we present an approach to solution of individual tasks along with methods proposed in the NAZOU research project. We evaluated methods for particular tasks and experimented with their collaboration in accomplishing the whole task using the developed software tools that are integrated using our framework for creation of adaptive portal solutions. Detailed information is presented elsewhere and summarized in two volumes of NAZOU research project workshop proceedings [17]. Fig. 3 presents variability of proposed methods and their corresponding tools developed.

4.1 Acquisition of documents and data from the Web

We proposed three approaches to acquisition of documents (job offers for us) and data (relevant parts of the documents) from the Web: (i) manual acquisition, (ii) automatic acquisition by browsing the Web, and (iii) automatic acquisition by downloading data from known sites that provide required information.

In case of manual acquisition of offers there is a human who fills the information base of offers. We developed special editor JOE (Job Offer Editor), which enables the providers of job offers to insert offers in such a manner that it is represented by an ontology. The point is that the huge space of the Web has to be narrowed to the particular domain. We seek a representation, which will retain all the essential information. Moreover, as we already mentioned, it will even enrich the data extracted from documents acquired from the Web by using methods of semantic annotation [14]. Annotation is a difficult task as we require machines to find information, which is often evident to a human, but definitely not to a machine [18]. Manual insertion of offers is important especially for experimenting with methods for information processing, as in such a case we can check the input data and relate them to the expected results while organizing and presenting offers [4].

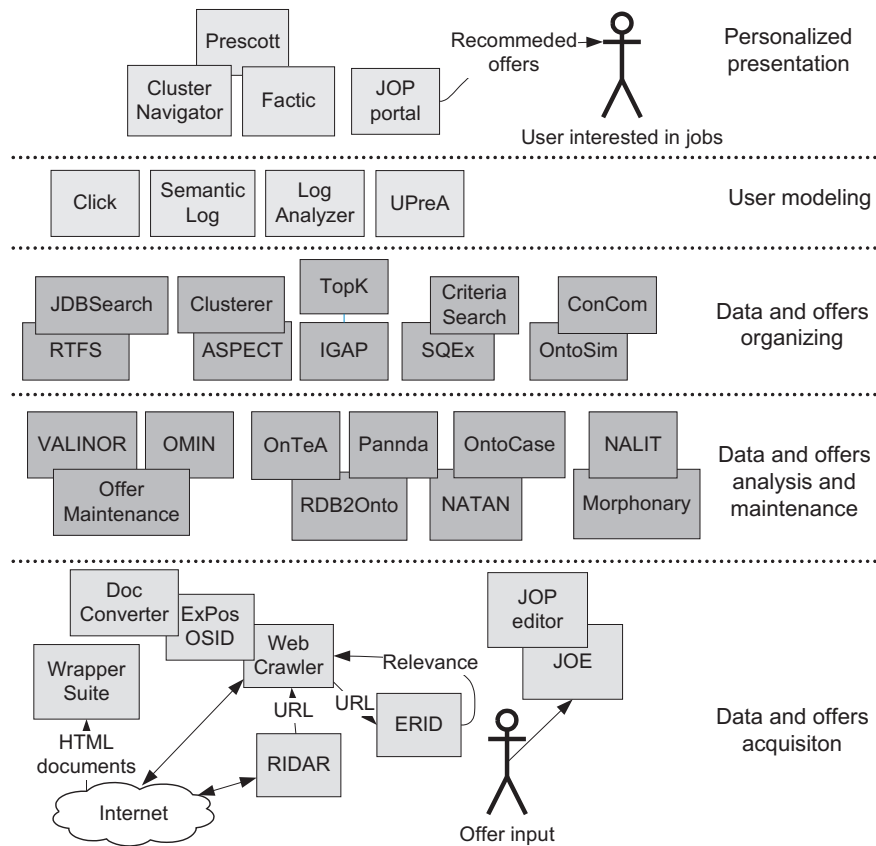


Fig. 3. Tools for acquisition, analysis, organization, and presentation of data and offers.

Automatic acquisition of documents by browsing the Web is based on the concept of focused crawling. In the first step, pages that definitely do not contain offers are filtered out. Consequently the offer is extracted from the web page and it is saved into the corporate memory of the information system. We proposed a method of downloading offers realized by chain of several tools: WebCrawler, OSID – Offer Separation for Internet Documents, RIDAR – Estimate Relevance for Internet Documents, ExPoS – Job Offer Extraction from the Web Page [10].

Automatic acquisition of documents by a simple downloading from the known sites is suitable when we know where the information of given type is present (in our case job offers concentrated in several known web portals). This approach is based on constructing wrappers – software tools that extract the required content knowing the structure of the web pages. Important issue in this approach is design of methods for effective creation of wrappers. This includes methods of learning from positive and negative examples so that the creation of a wrapper (which to some extent still requires manual intervention) was effective and so it was able to react to the structure changes of the pages [8].

4.2 Analysis and organization of data and information

In the contemporary Web we find data that provide information and knowledge to humans, however for automatic processing we need to enrich data with its semantics and also with other information needed for effective processing (e.g., indexes important for browsing). There exist several ways how to analyze and organize data so that we get meaningful information. We focus on some aspects that we consider from a certain point of view as representative for this area. They are methods which aim at:

- *annotating and reasoning*: methods serve to enrich acquired data with additional information and meaning [14]; results are used in search when significant information in acquired documents is discovered (e.g., name of the company), or during the presentation by supplementing additional information (e.g., about geographical position of the place) or supplementing relationships between offers important for effective navigation;
- *fulltext indexing*: methods serve to support fulltext browsing in information space (indexed job offers) according to assigned criteria;
- *clustering*: methods for grouping data based on selected criteria; identification of similarity in information space is important for categorization and recommendation, for example in case that the user shows interest in some information (presented offer) we recommend him also similar offers [2];
- *searching*: methods serve to search in information space; besides standard keyword based methods we consider such methods that make use of intelligent query expansion based on an estimation of the user interest,
- *categorization*: methods serve to arrange data according to various criteria; e.g. methods for creating ordered lists and generating rules for classification. As an example we mention a method for induction of regulations for monotonous classification of job offers through which we obtain top-k items concerning assigned preferences of the user [13];
- *text processing*: methods serve for the above mentioned tasks where text analysis is often requested, for example for comparison of offers or their annotation.²

4.3 Information presentation

When designing methods for presentation we focused on enriching the information space with adaptation to the user and his context. We especially concentrated on adaptive presentation of the content and adaptive navigation in hyperspace [2]. Our goal was to present information to the user in a personalized fashion, i.e. information, which is relevant for him and in addition in

² Note that in the field of natural language processing there is a major disproportion between the achieved progress for processing English and other languages (including Slovak). It is given by the languages alone (analysis of a flexive language is more difficult than of English), but also by the volume of applications and effort including means spend on text processing in respective language.

such fashion which best suits his needs [9]. For this we proposed the method of user behavior analysis, which comes out from defined heuristics with regard to "clicks" of the user (e.g., the meaning of the first activities of the user while browsing information space is provably higher for stating interests of the user than the other ones) [3].

For presentation itself it is possible to use several approaches. It should be noted that we deal with problem of ontology visualization. We proposed two views – one is based on facets that serve for navigation by constraining the information space and the second on visual navigation in clusters [21].

5 Conclusion

The Web contains information about a large number of questions, which can be of interest for us already today – and its content grows day by day. It is becoming one of the most important sources of information, it is just necessary to know how to obtain them from it. Concerning the scale and other properties of the Web, this is not at all a simple task. Without suitable tools the absolute majority of information would stay hidden for the user.

A software tool is an outcome of a development that must be preceded by research of the Web itself, by researching for new methods of data acquisition, organization, and presentation. This research, as this paper tried to outline, has already brought results, but it is clear the research and development must go on. We need to get to know the Web better. It is obvious that it is developing and thus changing constantly. This by no means makes investigating it easier. A change that could make acquiring information easier is to enrich what is written on the Web with at least some indication of its meaning (semantics). This opens room for research of methods that could be more fundamentally different from what we know today.

This work was partially supported by the State programme of research and development, SPVV1025/04, by the Slovak Scientific Grant Agency, VG1/0508/09) and by the Slovak Research and Development Agency, APVV-0391-06.

References

1. Ahmad, M.N., Colomb, R.M.: Managing ontologies: a comparative study of ontology servers. In: J. Bailey, A. Fekete (eds.), Proc. of the Conf. on Australasian Database, ACM Press, (2007), pp.13-22.
2. Andrejko, A., Bielíková, M.: Comparing instances of ontological concepts for personalized recommendation in large information spaces. Computing and Informatics, 2009, to appear.
3. Barla, M., Tvarožek, M., Bielíková, M.: Rule-based user characteristics acquisition from logs with semantics for personalized web-based systems: Updates of logic programs. Computing and Informatics, 2009, to appear.

4. Bartalos, P. et al.: Building an Ontological Base for Experimental Evaluation of Semantic Web Applications. In: Jan van Leeuwen et al. (eds.), LNCS 4362, Sofsem 2007, Springer (2007), pp. 682-692.
5. Brusilovsky, P., Kobsa, A., Nejdl, W.(eds.): The Adaptive Web: Methods and Strategies of Web Personalization, LNCS 4321, Inf. Sys. and Appl., Springer (2007).
6. Brusa, G., Caliusco, M. L., Chiotti, O.: A process for building a domain ontology: an experience in developing a government budgetary ontology. In: M.A. Orgun et al. (eds.), Proc. of Workshop on Advances in Ontologies, ACM Press, (2006), 7-15.
7. Ding, L., Kolari, P., Ding, Z., Avancha, S.: Using Ontologies in the Semantic Web: A Survey. In: Sharman, R. et al. (eds.), Ontologies: A Handbook of Principles, Concepts and Applications in Information Systems, Springer, (2007), pp. 79-113.
8. Frivolt, G., Kisac, I.: Interactive Wrapper Learning for Automatic Data Gathering. In: Proc. of Tools for Acquisition, Organisation and Presenting of Information and Knowledge (2), Research Project Workshop, (2007), pp. 63-67.
9. Gurský, P., Horváth, T., Jirašek, J., Krajčí, S., Novotný, R., Pribolová, J., Vaneková, V., Vojtáš, P.: User preference web search experiments with a system connecting web and user. Computing and Informatics, 2009, to appear.
10. 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, Workshop on Theory and Practice of IT, (2005), pp. 41-46.
11. Heckmann, D., Schwartz, T., Brandherm, B., Schmitz, M., von Wilamowitz-Moellendor, M.: Gumo – the general user model ontology. In L. Ardissono et al. (eds), Int. Conf. on User Modeling, Springer, LNCS 3538 (2005), pp. 428-432.
12. Hoang, H.H. et al.: Towards a New Approach for Information Retrieval in the SemanticLIFE Digital Memory Framework. In: IEEE/WIC/ACM Int. Conf. on Web intelligence. IEEE CS Press, (2006), pp. 485-488.
13. Horváth, T., Vojtáš, P.: Ordinal Classification with Monotonicity Constraints. In: Proc. of the 6th Industrial Conf. on Data Mining, LNAI 4065, Springer (2006), pp. 217-225.
14. Laclavík, M., Šeleng, M., Gatial, E., Hluchý, L.: Ontea: Platform for Pattern based Automated Semantic Annotation. Computing and informatics, 2009, to appear.
15. Machová, K., Bednár, P., Mach, M.: Various Approaches to Web Information Processing. In Computing and Informatics, Vol. 26, 2007, No. 3, pp. 301-327.
16. Návrat, P., Bieliková, M., Rozinajová, V.: Acquiring, Organising and Presenting Information and Knowledge from the Web. In B. Rachev, A. Smrikarov (eds.), Proc. of CompSysTech'06, , Bulgaria (2006).
17. Návrat, P., Bartoš, P., Bieliková, M., Hluchý, L., Vojtáš, P.: Tools for Acquisition, Organisation and Presenting of Information and Knowledge, Research Project Workshop, Proceedings, (2006, 2007).
18. Nekvasil, M., Svátek, V., Labský, M.: Transforming Existing Knowledge Models to Information Extraction Ontologies. In: 11th Int. Conf. (BIS'08), Springer (2008), pp.106-117.
19. Oberle, D. et al.: DOLCE ergo SUMO: On Foundational and Domain Models in SWIntO. In J. of Web Semantics, 5 (2007) 156-174.
20. Shadbolt, N., Hall, W., Berners-Lee, T.: The Semantic Web Revisited. IEEE Intelligent Systems, (May/June 2006), 96-101.
21. Tvarožek, M., Bieliková, M.: Visualization of Personalized Faceted Browser Interfaces. In: P. Forbrig et al. (eds.), IFIP Int. Federation for Information Processing, Vol. 272, Human-Computer Interaction Symposium, Springer (2008), pp. 213-218.