Methods and Tools for Acquiring and Presenting Information and Knowledge in the Web

Pavol Návrat, Mária Bieliková, Viera Rozinajová

Abstract: *Aiming at a qualitatively new model of processing information in a heterogeneous environment such as the current and the future Web, it is desirable to focus among others on investigating new ways of working with information and knowledge in a heterogeneous environment, especially with imperfect and vague information from perhaps dubious sources. They should include acquiring information and knowledge, organising information and knowledge, including categorisation based on ontologies, presenting information and knowledge in a way appropriate for a human, verifying, measuring and determining quality of information and knowledge (relevancy and trustworthiness in particular), designing models, methods and techniques needed for an efficient processing of information and knowledge in an environment of heterogeneous sources. Research such as the outlined one should involve verification of results by means of extensive experimentation with one or several pilot applications in selected domains of application.*

Key words: *Information, Knowledge, Semantic Web, Ontology.*

INTRODUCTION

Amount of accessible information and knowledge grows at an unprecedented pace. Its extent, quality and accessibility changes also due to the world-wide use of the Internet. The Internet and its services (e.g. World Wide Web or electronic mail) can be used as a very appropriate environment for the research of new ways of knowledge acquisition from the heterogeneous sources, organizing, validation, evaluation and maintenance of actual knowledge. Knowledge originates in information, which in turn is just data interpreted in a certain context. Internet forms a distributed environment for the heterogeneous sources of information. The distribution of information is important from the point of view of information accessibility, while heterogeneity is the key feature of the documents represented and presented on the Internet.

When looking for a certain information we are often overwhelmed by a huge amount of data of various kind and quality. For instance, in the case of e-mail, an effective filtering of messages would be very helpful. In this sense issue of reliability of the information sources becomes very important. Many search tools provide too extensive and irrelevant answers to user queries. On the other side, these tools are not able to provide information which is on the Internet, but is represented in a form that is difficult to process. In principle, search tools and services can be divided into two groups: not specialised - their main concern is to search, index and organise potentially all the sources found on the Internet. Specialised tools and services focus on a certain area of interest and present such information. The next step in the development of the search tools are so-called meta-search tools and services which integrate the results of more basic search tools. Examples of such systems are Meta Crawler, DogPile or ProFusion.

Some of the search tools seek to present information according to the needs of certain users. They allow the user to adapt some features of the presented information creating in this way an adaptable presentation. When using more sophisticated way of adapting to the user's needs, an adaptive presentation can be created [4], [5]. This means that the relevant information is presented according to the user model. The user model represents knowledge about the user, his preferences, and characteristics. It can also include the usage data, i.e. data about the user's navigation in the hyperspace.

Next layer to be added could be a learning module and a module of automatic acquisition of the knowledge about the user – i.e. his preferences, including their modelling.

Besides the awareness of the user model it also important to be aware of the main features of an environment, in which the user works [3], [7]. The knowledge about the environment is represented in a context model. Results of the search process can be influenced for instance by the type of the end device (PC, PDA or mobile phone). Also "social" aspects of the environment could play an important role (whether the user is communicating with someone else, the time of presenting information, etc.)

To achieve the main goal – i.e. to provide up-to-date information, another important ability to be researched is the ability to learn behaviour of information sources and monitor them accordingly, allowing for an efficient retrieval of most up-to-date information [14], [13].

Further improvement of the search process could be achieved by specifying the queries more precisely (for instance by removing key words in Alta Vista), by improving of indexing, by resorting the found sources, by collaborative filtering enriched with content-based analysis or by taking into account not only the content of the information sources but also interconnection of them (using HITS or PageRank algorithm) [10], [18], [19], [20]. Another approach to information acquisition is the navigation support in the Web hyperspace (for instance marking recommended references in the presented documents in the system Personal WebWatcher), which can be further improved by support of its modelling [8]. A further improvement could be expected if the query would be augmented by words from external vocabulary (thesaurus, semantic network, dictionary of synonyms) [17].

One interesting way of knowledge acquisition is visualization of relations among data by linear or nonlinear projection of three-dimensional graphs representing data. Such a visualisation can also lead to a discovery of data patterns.

One problem of the methods used nowadays in the information processing is the fact that references to billions of on-line documents are not effective enough to navigate to relevant information. Usually too many documents are found and it is very difficult to find the right ones. The search tools support searching the given words in the content, i.e. they facilitate finding documents which contain these words or phrases. But often the situation is more complicated: one looks for the information about "something", but the relevant documents need not always contain the given word. Therefore the goal is to "rebuild" Internet and the Web to the network of information that could be processed by machines, as opposed to the present state where information is understandable only by people. The Web is developing from a weakly structured network towards a more structured network enriched by data that help reading or inferring meaning (semantics).

TOWARDS THE SEMANTIC WEB

Transforming the Web towards the semantic Web encompasses a gradual transformation of representation of the existing documents to a representation which enriches the presented information by semantic concepts that can be utilised in its automated processing [6]. Currently, the basis are mark-up languages, which have been instrumental in the extensive development of the present Web. XML defines simple but strictly imposed syntactic rules. It forms a syntactic basis for defining mark-up languages for expressing various kinds of documents. Moreover, there are methods and tools available allowing to include metadata and document validation, making use of the XML Schema. However, XML is not a language to express meaning of the structured data being represented. This is where the RDF (Resource Description Framework) comes in. It allows defining (semantic) relationships between objects on the Web. Objects are identified by URI. Relationships are denoted by concepts defined in some ontology (a dictionary of concepts). Mutual "understandability" of attributes and their values is secured by creating and reusing of definitions of concepts in ontologies in languages such as OWL.

Acquiring, organizing and maintaining knowledge from the Web can be implemented in a distributed way by heterogeneous autonomous agents [1], [16]. Agents equipped with knowledge are able to search databases but also the Internet. Some of the open problems relate to models of collaboration in a heterogeneous environment such as the Internet.

Software systems in general – and particularly those in the domain of hypermedia – are developed still mostly without some organised reuse. Instead of an ad hoc reuse, we advocate for a development for reuse, which is a basis of domain-oriented approaches to software development [21], [22].

METHODS AND TOOLS

Aiming at a qualitatively new model of processing information in a heterogeneous environment such as the current and the future Web, it is desirable to focus among others on investigating new ways of working with information and knowledge in a heterogeneous environment such as the Web, especially with imperfect and vague information from perhaps dubious sources:

- acquiring information and knowledge,
- organising information and knowledge, including categorisation based on ontologies,
- presenting information and knowledge in a way appropriate for a human,
- verifying, measuring and determining quality of information and knowledge (relevancy and trustworthiness in particular),
- designing models, methods and techniques needed for an efficient processing of information and knowledge in an environment of heterogeneous sources.

Research such as the outlined one should involve verification of results by means of extensive experimentation with one or several pilot applications in selected domains of application.

In the first phase, inevitably the research should focus on investigating models of a heterogeneous environment. Basis is the structure of the net (nodes and links between them), which allows to make use of the results of research in the area of models of hypermedia systems (Dexter model, enriched with adaptive features in the Munich model, and others) [3]. When modelling hyperspace of the heterogeneous environment, it is desirable to consider also a time dimension, i.e. existence of versions of particular sources of information, and also a variant dimension, i.e. alternative contents [2]. Such a model can then serve for navigation in the hyperspace. As an example of variants, which is potentially relevant for many Web sources, are language versions of the contents of the information sources.

The model should include also data about information and knowledge incorporated in the hyperspace of the heterogeneous space. In particular, research should focus on using metadata that express semantics of the contents. Initial point are the standards as defined by the W3C, enriching the Web with semantics. These standards allow defining ontologies, which form a shared terminology within a description of contents of the information sources. Ontologies in turn define meaning of the contents by means of dictionaries of concepts and mechanisms of inference over them. Inference opens door for a better utilisation of the contents by software tools. Research should seek methods of ontology design and finding relations between similar ontologies. Software tools should be devised that would effectively use the metadata when processing the contents of the information sources.

Once we have a model of a domain expressed using an ontology, it is possible to devise searching and reasoning mechanisms that would produce knowledge to the end user that was altogether not obvious to yield. Involving ontology in searching allows sorting

data not only according to key words (as most of the search machines do), but allows also sorting according to inferred properties and found connections.

Acquiring knowledge from ontological models can be defined in a very specialised way [9]. There are, however, also some common techniques for finding relations between information chunks, which are based on finding a similar case (Case Based Reasoning) or finding a statistically similar case (Rule Based Reasoning).

As far as formal models of description of semantics are concerned, various systems of working with an imprecise information (logical and probabilistic ones) and relations between them should be investigated [15].

At the heart of any method or tool, there are methods and techniques for searching information. A user query usually contains only a few terms, which only insufficiently reflect the intended meaning of the query. We see a possibility of improvement in enriching queries by using external dictionaries. To maintain (or improve) precision, the enriching terms should be restricted to the domain of query.

Currently, most of the indexed pages are still plain HTML pages. However, they constitute only a fraction of information actually available on the Web, since most part is hidden in databases of information sources that are exported into HTML but cannot be easily indexed (and thus searched). Also, it is desirable to investigate a distributed searching within several independent sources.

One of the possible ways of acquiring information and knowledge is by visualisation of relations between data. It can have a form of linear or non-linear projections of three dimensional graphs that represent data with a monographic or stereographic mapping using systems such as Web3D or an embedded virtual reality.

Retrieved information should be integrated, organised and sorted so that its presentation is simpler. Here, methods of categorisation and clustering will be helpful [11]. In particular, methods of mathematical statistics, neural networks, evolution algorithms and probabilistic models should be investigated.

We see as an important dimension of any research in this area investigating possibilities of adaptation in reference to all the basic activities, i.e. to knowledge acquiring, organising and also presenting to the user. Crucial is adaptive presentation of information and adaptive navigation in the hyperspace. The goal is to devise methods and tools that would allow presenting to the users personalised information, i.e. such that is relevant to the user and also presented in a way that suits best to each particular user. This may require creating a suitable structure for representation of knowledge on the user (her or his characteristics, preferences, level of knowledge) – a user model. When adopting a presentation, options of taking into account a context of the presentation (e.g. location where the user is, and the device she is using).

User models are to be used in automated recommending of actual and relevant information. One of the approaches is recommending information according to the textual contents of the information. Here, employing ontologies and semantic analysis of words seems to be essential. Alternatively, recommending information can be accomplished by collaborative filtering, where preferences of people who are more or less regular consumers of information are used to determine recommendations to users with similar interests. Besides simpler heuristic methods, there are more sophisticated methods based on probabilistic models of preferences.

CONCLUSIONS AND FUTURE WORK

Results of any research should be justified at least in some context. We work on two pilot applications which are proposed so that they will allow testing extensively methods and tools for acquiring information. We believe to devise ways and tools for acquiring, organising and maintaining information and knowledge as well as presenting it so that it is suitable both for the human user and automatic processing. As a result, services will be

provided by Internet which will offer to user a better quality of information and knowledge (flexibility, trustworthiness, relevance etc.).

First pilot application we work on is in the domain of labour market. The intention is to provide information and knowledge on job offers. It is a typically heterogeneous problem in several dimensions. The sources themselves, containing job offers, can be very different – from multinational concerns to small local enterprises. Way of job offer presentation will also be different, although some common features in their structure and contents could be discovered. Language of presentation can also be different – just in the European Union, we have more than a dozen of official languages, not to speak of others. Various professional groups can have their specific concepts used in formulating job offers. Regional variations are also possible. Within all such cobweb of manifold heterogeneity it is necessary to retrieve and present to the user information which presents the most precise recommendations reflecting her or his needs or preferences. The provided information must be up to date. Our tools should contribute to a better access to information that would eventually help the user find a job she or he seeks.

Similar applications can be specified in other domains such as

- continuously updated electronic newspaper,
- active and passive tourism opportunities,
- selling, buying and renting immobility,
- study materials on some topic,
- latest results of research and development in some area.

Our tools should be generic enough to allow developing these and similar applications.

ACKNOWLEDGEMENTS

This work was supported by Science and Technology Assistance Agency under the contract No. APVT-20-007104 and by Scientific Grant Agency of Slovak Republic grant No. VG1/0162/03.

REFERENCES

[1] Balogh, Z., Laclavík, M., Hluchý, L., Budínska, I., Krawczyk, K. REMARK - Reusable Agent-Based Experience Management and Recommender Framework In: Proc.of International Conference on Computational Science, Part III, Krakow, LNCS 3038, Springer-Verlag, 2004, pp. 599-606.

[2] Bieliková, M., Návrat, P. Modelling versioned hypertext documents. In B. Magnusson, editor, System Configuration Management, ECOOP'98 SCM-8 Symposium, pages 188-197, Brussels, Belgium, July 1998. Springer-Verlag, LNCS 1439.

[3] Bieliková, M. Adaptive hypermedia presentation on the Web. In: L. Popelínský (Ed.): *Proc. of DATAKON 2003*, Brno 2003, pp. 72-91.

[4] Brusilovsky, P. Adaptive Hypermedia. *User* Modeling and User-Adapted Interaction, Kluwer academic publishers, 1996, 11:1-2, pp. 87-110.

[5] Brusilovsky, P. Methods and Techniques of Adaptive Hypermedia. User Modeling and User-Adapted Interaction, Kluwer Academic Publishers, 2001, 6:2-3, pp. 87-129.

[6] Davies, J., Fensel, D., van Harmelen, F. Towards to Semantic Web: Ontology-driven Knowledge Management, John Wiley & Sons, 2003.

[7] Dolog, P., Bieliková, M. Hypermedia Systems Modelling Framework. Computing and Informatics, 2002, 21:3, pp. 221-239.

[8] Dolog, P., Bieliková, M. Navigation Modelling in Adaptive Hypermedia. In: P. De Bra, P. Brusilovsky, R. Conejo (Eds.): Proc. of Int. Conference on Adaptive Hypermedia – AH 2002, Springer Verlag, LNCS 2347, Málaga Spain, 2002, pp. 586-591.

[9] Filkorn, R., Návrat, P. Feature-based Filtering in Semantic Web. In: B. Thalheim and G. Fiedler(Eds.): Emerging Database Research in East Europe, Proceedings of the Pre-Conference Workshop of VLDB 2003, Computer Science Reports, Brandenburg University of Technology at Cottbus, Report 14/03, pp. 46-50.

[10] Gurský, P., Horváth, T. Dynamic search of relevant information. In: Proc. Znalosti 2005, pp. 194-201.

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

[12] Jenčušová, E., Jirásek, J. Formal methods of Analysis of Security Protocols, Tatra Mountains Math. Publ. 25 (2002), pp. 1-10.

[13] Koval, R., Návrat, P. Intelligent Support for Information Retrieval of Web Documents. Computing and Informatics, 21, 5,2002, pp. 509-528.

[14] Koval, R., Návrat, P. Intelligent Support for Information Retrieval in the WWW Environment. In: Proc. of ADBIS 2002 - Advances in Databases and Information Systems, Manolopoulos, Y. and Návrat, P. (Eds.), Springer LNCS 2435, pp 51-64, 2002.

[15] Krajči, S., Lencses, R., Medina, J., Ojeda, M., Vojtáš, P. A similarity based unification model for flexible querying. In: T. Andreasen et al eds.: Proc. FQAS'02, LNCS 2522, Springer Verlag, Berlin 2002, pp. 263-273.

[16] Laclavík, M., Balogh, Z., Hluchý, L., Slota, R., Krawczyk, K., Dziewierz, M. Distributed Knowledge Management based on Software Agents and Ontology. In: R.Wyrzykowski et.al. eds. Proc. of 5-th Intl. Conf. on Parallel Processing and Applied Mathematics PPAM'2003, LNCS 3019, Springer-Verlag, 2003, pp. 694-699.

[17] Lencses, R. Indexing for Information Retrieval System supported with Relational Database, In: Proc. Sofsem 2005, Slovakia, pp. 81-90, 2005.

[18] Pokorný, J., Vojtáš, P. A data model for flexible querying. In A. Caplinskas and J. Eder eds.: Proc. ADBIS'01, LNCS 2151, Springer Verlag, Berlin 2001, pp. 280-293.

[19] 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.

[20] Polčicová, G., Slovák, R., and Návrat, P. Combining Content-based and Collaborative Filtering. In: Masunaga, Y., Pokorný, J., Štuller, J., and Thalheim, B.: Proceedings of Challenges, ADBIS-DASFAA Symposium on Advances in Databases and Infromation Systems 2000, 118-127, 2000.

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

[22] Vranić, V. Multi-paradigm design with feature modeling. Computer Science and Information Systems Journal (ComSIS). Accepted for publishing (2005).

ABOUT THE AUTHORS

Prof. Pavol Návrat, PhD. navrat@fiit.stuba.sk, Assoc. Prof. Mária Bieliková, PhD., bielik@fiit.stuba.sk, Viera Rozinajová, PhD., rozinajova@fiit.stuba.sk, all from:
Institute of Informatics and Software Engineering, Faculty of Informatics and Information Technologies, Slovak University of Technology in Bratislava, Slovakia.