Connecting Web and User

Jan Dědek[1], Alan Eckhardt[2], and Peter Vojtáš[2]

[1] Charles University, Faculty of Mathematics and Physics, Prague, Czech Republic
[2] Academy of Sciences of the Czech Republic, Institute of Computer Science

Abstract. The paper summarizes our research during last two years. We are concentrated on the problem of connecting web and user. This problem consists of two main aspects: user preference modelling and web content mining. We see the modelling of uncertainty beneficial to both problems and we have invested indispensable effort to uncertainty issues of our solutions. We see the solution of our problem (connecting web and user) in the most recent idea of web semantization, which will be also presented in this paper.

1 Introduction

This paper summarizes our research during last two years. We are concentrated on the problem of connecting web and user. This problem consists of two main aspects: web content mining and user preference modelling.

Web content mining is supposed to extract structured information from possibly heterogeneous web resources. From known structure of the extracted information we can easily deduce semantics of the information and such information can be further used for precise semantic information querying. This principle is widely developed in the idea of the Semantic web. We have experimented with web content mining and exploited two different approaches: HTML structural induction and linguistic analysis. More details are presented in section 2.

Combination of web content mining and the idea of Semantic web leaded into the formulation of the idea of gradual *web semantization*, which is described in the section 5.

Modelling of user preferences helps user to find the most interesting information, products, offers, service, etc according to his or her preferences. Modelling of user preferences in the background of Semantic web is even more interesting and can bring useful solutions. More details about our work in this field are presented in section 3.

We see both problems (web content mining and user preference modelling) very difficult and the results are usually uncertain because they are influenced by human factor. So we see the modelling of uncertainty beneficial to both problems and we have invested indispensable effort to uncertainty issues of our solutions. More details are presented in section 4.

The whole situation is presented in the figure 1

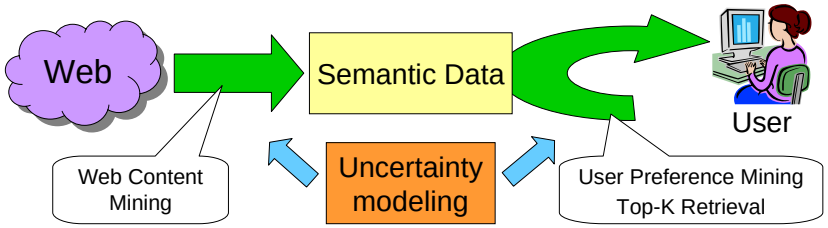

Fig. 1. Connecting web and user.

2 Web Content Mining

Web content mining or web information extraction splits pages to dominantly tabular and/or textual. We will discuss our experience with both separately.

2.1 Dominantly Tabular Pages

In our works [1, 2] we have developed a software web information extraction tool that utilizes repeating structural patterns present on "product summary" HTML web pages. This tool reduces human work need to extract information from such web pages. In the present this tool can be used with advantage, but is still in development and promise better results with future new features (reduction of effort connected with definition of regular expressions and exploitation of "detail pages" in internet shop sites).

2.2 Dominantly Textual Pages

In [3] we have introduced a linguistic-based method for extraction of information from text-based web resources in Czech. We have experimented with several linguistic tools for Czech, namely Tools for machine annotation – PDT 2.0 and the Czech WordNet. Then in [4] we have presented a design of a system which captures text of web-pages, annotates it linguistically by PDT tools, extracts data and stores the data in an ontology. We reported on initial experiments in the domain of reports of traffic accidents. The results showed that this method can e.g. aid summarization of the number of injured people.

In [5] we focused on the data extraction phase of the system and presented methods for learning queries over linguistically annotated data. Machine learning procedure for automated learning of the linguistic queries (extraction patterns) was presented in [6] where we have used ILP as the learning method. Semantic interpretation of extraction patterns and subsequent semantic interpretation of extracted data were developed in [7].

3 Modelling of User Preferences

User preferences became recently a hot topic. The massive use of internet shops and social webs require the presence of a user modelling, which helps users to

orient themselves on a page. There are many different approaches to model user preferences. In [8] we made overview of the state-of-the-art in the area of acquisition of user preferences and their induction. Main focus was on the models of user preferences and on the induction of these models, but also the process of extracting preferences from the user behaviour was studied. We also presented our contribution to the probabilistic user models in [8].

We focused on models of user preferences in Semantic web in [9]. We presented a model for querying over RDF data with user preferences and for ordering of results by a user's aggregation function. This model has theoretical base in a modification of fuzzy description logic, which is embeddable into the two valued description logic which extends OWL. We described first experiments made with Tokaf - an implementation of framework for the flexible querying.

Models of user and group preferences in social networks and the Semantic web were discussed in [10]. We have constructed a model for user and group preference querying over RDF data as well as for ordering of answers by aggregation of particular attribute ranking.

In [11] we generalized Fagin's algorithm for getting top-k answers according to user preferences. The generalization was made in two directions – we developed some new heuristics for top-k search in the model without random access and proposed a method of ordering lists of objects by user fuzzy function. To enable different user preferences our system does not require objects to be sorted – instead we use a B+ tree on each of the attribute domains. This leads to a more realistic model of Web services.

Web search heuristics based on Fagin's threshold algorithm assume we have the user profile in the form of particular attribute ordering and a fuzzy aggregation function representing the user combining function. Having these, there are sufficient algorithms for searching top-k answers. Finding particular attribute ordering and aggregation for a user still remains a problem. In [12] our main contribution is a proof of concept of a new *iterative* process of acquisition of user preferences and attribute ordering.

Usually different users have different fuzzy scoring function – a user preference model. Main goal of [13] was to assign a user a preference model automatically. To achieve this we decomposed user's fuzzy ranking function to ordering of particular attributes and to a combination function. To solve the problem of automatic assignment of user model we design two algorithms, one for learning user preference on particular attribute and second for learning the combination function. Methods were integrated into a Fagin-like top-k querying system with some new heuristics and tested.

In connection with the development of the idea of web semantization we proposed a decomposed model: semantic search engine – user preference agent, which was presented in [14] and [15].

In [16] we deal with the problem of learning user preferences from user's scoring of a small sample of objects with labels from a very small linearly ordered set to use these preferences for a top-k query. We merged our terminology with economics and identified fuzzy ranking function to ordering of particular attributes

as an *objective function* and our combination function as an *utility function*. One of the main contributions of this last work ([16]) was that we have compared our method to some classical data-mining methods. We used several measures (RMSE and rank correlations ...) to evaluate efficiency of these methods.

4 Modelling of Uncertainty

We are interested in replacing human processing of web resources by automated processing. Based on an experimental system we identified uncertainty issues making this process difficult for automated processing and tried to minimize human intervention in [1] and [2]. In particular we focus on uncertainty issues in a web content mining system and a user preference mining system.

In [17] we discussed the what, who, when, where, why and how of uncertain reasoning based on achievements of URW3XG[3], our experiments and some future plans.

What and Why – improving semantic web practice through uncertain reasoning.

Who and When – will create, maintain and use this annotation. Will this annotation be done by a human creator using an annotation supporting tool for web page creation? Or will it be done by a third party annotation? For this, we discussed a refinement of URW3XG use cases. Possible use of this enriched web will be for humans and services.

Where – will be this annotations stored. Our proposal is based on the web crawler Egothor repository[4] (we have crawled data in size of several TB from .cz domain) and an additional semantic repository build on the top using data pile technology [18].

How – to semantically enrich information and how to measure success and/or progress of such enrichment. This problem consists of two parts, namely, a data mining task and an ontology modelling task. Third party annotation of great size can be done only in an automated way and it should be done according to an ontology.

In [17] and [15] our annotation ontology grew out of URW3XG uncertainty ontology and extended some features needed for annotation. We started here from an assumption that a part of annotation will be done by a web information extraction and that this is the main source of uncertainty.

Success of this approach can be measured primarily by the advance of semantic web functionalities. This is easier to measure for software agents. More difficult is to design metrics to measure human user satisfaction. All these aspects were discussed in the presentation of [17].

[3] `http://www.w3.org/2005/Incubator/urw3/XGR-urw3/`
[4] `http://www.egothor.org/`

5 Web Semantization

Nobody seems to care in the semantic web community about the content of the web of today or of pages published without annotations. By our opinion the content of the web of today is too valuable to be lost for emerging semantic web applications. The problem of *semantization* (enrichment) of current web content as an automated process of third party annotation for making (at least a part, increasing in time) of today web accessible for machine processing and hence enabling intelligent tools for searching and recommending things on the web. Our main idea (presented in [15]) is to create a semantic repository of information automatically extracted from the web and corresponding annotated copies of (mainly textual) parts of web pages, and make it available to intelligent agents.

Acknowledgments
This work was partially supported by the Ministry of Education of the Czech Republic (grant MSM0021620838) and by Czech projects 1ET100300517 and 1ET100300419.

References

1. Eckhardt, A., Horváth, T., Maruščák, D., Novotný, R., Vojtáš, P.: Uncertainty issues in automating process connecting web and user. In da Costa, P.C.G., ed.: URSW '07 Uncertainty Reasoning for the Semantic Web - Volume 3, The 6th International Semantic Web Conference (2007) 97–108
2. Eckhardt, A., Horváth, T., Maruščák, D., Novotný, R., Vojtáš, P. In: Uncertainty Issues and Algorithms in Automating Process Connecting Web and User. Volume 5327 of Lecture Notes in Computer Science. Springer Verlag (2008)
3. Dědek, J., Vojtáš, P.: Extrakce informací z textově orientovaných zdrojů webu. In Snášel, V., ed.: Znalosti 2008. (2008) 331–334
4. Klopotek, M., Przepiórkowski, A., Wierzchoń, S., Trojanowski, K., eds.: Intelligent Information Systems XVI, Zakopane, Poland, Academic Publishing House EXIT (2008)
5. Dědek, J., Vojtáš, P.: Linguistic extraction for semantic annotation. In Badica, C., Mangioni, G., Carchiolo, V., Burdescu, D., eds.: 2nd International Symposium on Intelligent Distributed Computing. Volume 162 of Studies in Computational Intelligence., Catania, Italy, Springer-Verlag (2008) 85–94
6. Dědek, J., Eckhardt, A., Vojtáš, P.: Experiments with czech linguistic data and ILP. In Železný, F., Lavrač, N., eds.: ILP 2008 - Inductive Logic Programming (Late Breaking Papers), Prague, Czech Republic, Action M (2008) 20–25
7. Dědek, J., Vojtáš, P.: Computing aggregations from linguistic web resources: a case study in czech republic sector/traffic accidents. In Dini, C., ed.: Second International Conference on Advanced Engineering Computing and Applications in Sciences, IEEE Computer Society (2008) 7–12
8. Eckhardt, A.: Inductive models of user preferences for semantic web. In Pokorný, J., Snášel, V., Richta, K., eds.: DATESO 2007. Volume 235 of CEUR Workshop Proceedings., Matfyz Press, Praha (2007) 108–119

9. Eckhardt, A., Vojtáš, P.: Uživatelské preference při vyhledávání ve webovských zdrojích. In Dvorský, J., Krátký, M., Mikulecký, P., eds.: Znalosti 2007, VSB-TY Ostrava (2007) 179–190
10. Eckhardt, A., Pokorný, J., Vojtáš, P.: Integrating user and group preferences for top-k search. In Tjoa, A.M., Wagner, R.R., eds.: Database and Expert Systems Applications, Regensburg, Germany, IEEE (2007) 317–322
11. Eckhardt, A., Pokorný, J., Vojtáš, P.: A system recommending top-k objects for multiple users preferences. In Martin, T., ed.: 2007 IEEE Conference on Fuzzy Systems. IEEE Fuzzy systems, London, United Kingdom (2007) 1101–1106
12. Eckhardt, A., Horváth, T., Vojtáš, P.: PHASES: A user profile learning approach for web search. In Lin, T., Haas, L., Motwani, R., Broder, A., Ho, H., eds.: 2007 IEEE/WIC/ACM International Conference on Web Intelligence - WI 2007, IEEE (2007) 780–783
13. Eckhardt, A., Horváth, T., Vojtáš, P.: Learning different user profile annotated rules for fuzzy preference top-k querying. In Prade, H., Subrahmanian, V., eds.: International Conference on Scalable Uncertainty Management. Volume 4772 of Lecture Notes In Computer Science., Washington DC, USA, Springer Berlin / Heidelberg (2007) 116–130
14. Eckhardt, A.: Návrh agenta řízeného uživatelskými preferencemi. In Vojtáš, P., ed.: ITAT 2008 Informačné Technológie - Aplikácie a Teória, Hrebienok, Slovakia, September 2008, Košice, Slovensko (2008) 31–34
15. Dědek, J., Eckhardt, A., Vojtáš, P., Galamboš, L.: Sémantický web. In: DATAKON 2008, Brno (2008)
16. Eckhardt, A., Vojtáš, P.: Considering data-mining techniques in user preference learning. In: 2008 International Workshop on Web Information Retrieval Support Systems. (2008)
17. Dědek, J., Eckhardt, A., Galamboš, L., Vojtáš, P.: Discussion on uncertainty ontology for annotation and reasoning (a position paper). In da Costa, P.C.G., ed.: URSW '08 Uncertainty Reasoning for the Semantic Web - Volume 4, The 7th International Semantic Web Conference (2008)
18. Bednárek, D., Obdržálek, D., Yaghob, J., Zavoral, F.: Data integration using datapile structure. In Čaplinskas, A., Eder, J., eds.: Advances in Databases and Information Systems. LNCS 2151, Berlin Heidelberg, Springer Verlag (2005) 178–188