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 modeling and web content mining. Modeling of user preferences helps user to find the most interesting information, products, offers, service, etc.

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 modeling.

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. More details are presented in section 2.

Modeling of user preferences helps user to find the most interesting information, products, offers, service, etc according to his or her preferences. More details are presented in section 3.

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

The whole situation is presented in the figure 1

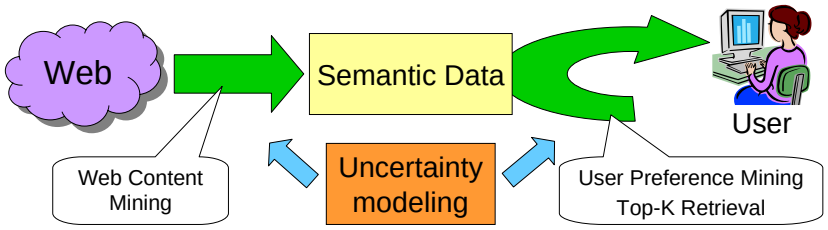

Fig. 1. Connecting Web and User

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

Uncertainty issues connected with information extraction (and annotation) from tabular pages were discussed in [1]. Extraction of textual pages will use techniques described in [6]. Both approaches (and any other approach) generate a level of (un)certainty they have about their annotations.

2.2 Dominantly Textual Pages

2.3 Uncertainty Issues and Algorithms in Automating Process Connecting Web and User

, Eckhardt, Alan Horváth, T. Maruščák, D. Novotný, R. Vojtáš, Peter, Uncertainty Reasoning for the Semantic Web I [1]

We focus on replacing human processing web resources by automated processing. On an experimental system we identify uncertainty issues making this process difficult for automated processing and try to minimize human intervention. In particular we focus on uncertainty issues in a Web content mining system and a user preference mining system. We conclude with possible future development heading to an extension of OWL with uncertainty features.

2.4 Uncertainty Issues in Automating Process Connecting Web and User

, Eckhardt, Alan Horváth, T. Maruščák, D. Novotný, R. Vojtáš, Peter, URSW '07 Uncertainty Reasoning for the Semantic Web - Volume 3 [2]

We are interested in replacing human processing of web resources by automated processing. Based on an experimental system we identify uncertainty issues making this process difficult for automated processing. We show these uncertainty issues are connected with Web content mining and user preference mining. We conclude with a discussion of possible future development heading to an extension of web modeling standards with uncertainty features.

2.5 Dedek, Vojtas: Znalosti2008

[3]

The authors present a linguistic-based method for extraction of information from text-based web resources. The paper deals with several linguistic tools for Czech, namely Tools for machine annotation – PDT 2.0 and The Czech WordNet.

2.6 Dedek, Vojtas: IIS 2008 (Zkopane)

[4]

The paper addresses a problem of information extraction from Czech texts from the Web. The method described in the paper exploits existing linguistic tools created originally for a syntactically annotated corpus, Prague Dependency Treebank (PDT 2.0). We propose a system which captures text of web-pages, annotates it linguistically by PDT tools, extracts data and stores the data in an ontology. We report on initial experiments in the domain of reports of traffic accidents. These experiments are promising, e.g. enabling summarization of the number of injured people.

2.7 Dedek, Vojtas: IDC 2008

[5]

Bottleneck for semantic web services is lack of semantically annotated information. We deal with linguistic information extraction from Czech texts from the Web for semantic annotation. The method described in the paper exploits existing linguistic tools created originally for a syntactically annotated corpus, Prague Dependency Treebank (PDT 2.0). We propose a system which captures text of web-pages, annotates it linguistically by PDT tools, extracts data and stores the data in an ontology. We focus on the third phase – data extraction – and present methods for learning queries over linguistically annotated data. Our experiments in the domain of reports of traffic accidents enable e.g. summarization of the number of injured people. This serves as a proof of concept of our solution. More experiments, for different queries and different domain are planned in the future. This will improve third party semantic annotation of web resources.

2.8 Dedek, Vojtas: ADVCOMP 2008 (Valencie)

[6]

Semantic computing aims to connect the intention of humans with computational content. We present a study of a problem of this type: extract information from large number of similar linguistic web resources to compute various aggregations (sum, average,...). In our motivating example we calculate the sum of injured people in traffic accidents in a certain period in a certain region. We restrict ourselves to pages written in Czech language. Our solution exploits existing linguistic tools created originally for a syntactically annotated corpus, Prague Dependency Treebank (PDT 2.0). We propose a solutions which learns tree queries to extract data from PDT2.0 annotations and transforms the data in an ontology. This method is not limited to Czech language and can be used with any structured linguistic representation. We present a proof of concept of our method. This enables to compute various aggregations over linguistic web resources.

2.9 Dedek, Eckhardt, Vojtas: ILP 2008

[7]

In this paper we present basic experiments that we have made in connection with our research in the domain of the Semantic Web. These experiments should demonstrate possibilities of employing ILP technique in the task of acquisition of semantic information from text of Czech Web pages. These experiments are preceded by complex linguistic analysis of the texts and the output of linguistic tools is processed in the ILP procedure.

3 Modeling of User Preferences

3.1 Inductive Models of User Preferences for Semantic Web

, DATESO 2007, Eckhardt, Alan [8]

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 them selfs on a page. There are many different approaches to model user preferences. In this paper, we will overview the current state-of-the-art in the area of acquisition of user preferences and their induction. Main focus will be on the models of user preferences and on the induction of these models, but also the process of extracting preferences from the user behaviour will be studied. We will also present our contribution to the probabilistic user models.

3.2 Uživatelské preference při vyhledávání ve webovských zdrojích

, Eckhardt, Alan Vojtáš, Peter, Znalosti 2007 [9]

This paper is focused on models of user preferences in semantic web. We present 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 a fuzzy description logic, which is embedable into the two valued description logic and which extends OWL. We describe experiments made with Tokaf - an implementation of framework for the flexible querying. We also test a new heuristic for the algorithm for searching top k answers.

3.3 Integrating user and group preferences for top-k search

, Eckhardt, Alan Pokorný, J. Vojtáš, Peter, Database and Expert Systems Applications (DEXA) 2007 [10]

We discuss models of user and group preferences in social networks and the Semantic web. We construct a model for user and group preference querying over RDF data as well as for ordering of answers by aggregation of particular attribute ranking. We have implemented our methods and heuristics into the Tokaf middleware framework prototype. We describe also experiments with Tokaf.

3.4 A system recommending top-k objects for multiple users preferences

, Eckhardt, Alan Pokorný, J. Vojtáš, Peter, 2007 IEEE Conference on Fuzzy Systems (FUZZ-IEEE) [11]

We discuss models of user preferences in Web environment. We construct a model for user preference querying over a number of data sources and ordering of answers by a combination of particular attribute rankings. We generalize Fagin's algorithm in two directions - we develop some new heuristics for top-k search in the model without random access and propose 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. We implement our methods and heuristics for search of top-k answers into Tokaf middleware framework prototype. We describe experiments with Tokaf and compare different performance measures with some other methods.

3.5 PHASES: A User Profile Learning Approach for Web Search

, Eckhardt, Alan Horváth, T. Vojtáš, Peter, 2007 IEEE/WIC/ACM International Conference on Web Intelligence - WI 2007 [12]

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 this short paper our main contribution is a proof of concept of a new iterative process of acquisition of user preferences and attribute ordering.

3.6 Learning different user profile annotated rules for fuzzy preference top-k querying

, Eckhardt, Alan Horváth, T. Vojtáš, Peter, International Conference on Scalable Uncertainty Management SUM 2007 [13]

Uncertainty querying of large data can be solved by providing top-k answers according to a user fuzzy ranking/scoring function. Usually different users have different fuzzy scoring function – a user preference model. Main goal of this paper is to assign a user a preference model automatically. To achieve this we decompose 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.

3.7 Návrh agenta řízeného uživatelskými preferencemi

, Eckhardt, Alan, ITAT 2008 Informačné Technológie - Aplikácie a Teória, Hrebienok, Slovakia, September 2008 [14]

Vize semantickeho webu vykresluje web tak, .ze bude pochopitelny pro stroje. Tento .clanek p.ristupuje k semantickemu webu z opa.cneho konce - od u.zivatele. Navrhneme softwaroveho agenta vyu.z.vaj.c.ho semanticka data z.skana anotac., ktery je bude prezentovat u.zivateli podle jeho preferenc.. Tento agent usnadn. u.zivateli hledan. a vyb.er idealn.ho objektu, podle jeho preferenc ..

3.8 Considering data-mining techniques in user preference learning

, Eckhardt, Alan Vojtáš, Peter, 2008 International Workshop on Web Information Retrieval Support Systems [15]

In this paper 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. The main task of this process is to use these preferences for a top-k query, which delivers the user with an ordered list of k highest ranked objects. We deal with a problem of many ties in the highest score. Two algorithms for learning objective and utility functions are presented. We experiment and compare them to some classical data-mining methods. We use several measures (RMSE and rank correlations . . .) to evaluate efficiency of these methods.

4 Modeling of Uncertainty

4.1 Uncertainty Issues in Automating Process Connecting Web and User

, Eckhardt, Alan Horváth, T. Maruščák, D. Novotný, R. Vojtáš, Peter, URSW '07 Uncertainty Reasoning for the Semantic Web - Volume 3 [2]

We are interested in replacing human processing of web resources by automated processing. Based on an experimental system we identify uncertainty issues making this process difficult for automated processing. We show these uncertainty issues are connected with Web content mining and user preference mining. We conclude with a discussion of possible future development heading to an extension of web modeling standards with uncertainty features.

4.2 Uncertainty Issues and Algorithms in Automating Process Connecting Web and User

, Eckhardt, Alan Horváth, T. Maruščák, D. Novotný, R. Vojtáš, Peter, Uncertainty Reasoning for the Semantic Web I [1]

We focus on replacing human processing web resources by automated processing. On an experimental system we identify uncertainty issues making this process difficult for automated processing and try to minimize human intervention. In particular we focus on uncertainty issues in a Web content mining system and a user preference mining system. We conclude with possible future development heading to an extension of OWL with uncertainty features.

4.3 Dedek, Eckhardt, Galambos, Vojtas: URSW 2008

[16]

In this position paper we discuss the what, who, when, where, why and how of uncertain reasoning based on achievements of URW3XG [2], our experiments and some future plans. What and Why - improving semantic web practice through uncertain reasoning. This vision is described in the URW3XG charter (see [2]), especially the objective is "to identify and describe situations [...] for which uncertainty reasoning would significantly increase the potential for extracting useful information; and to identify methodologies that can be applied to these situations and the fundamentals of a standardized representation that could serve as the basis for information exchange necessary for these methodologies to be effectively used." A crucial point in this is uncertainty annotation of web (extending W3C standards [3]). 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 will discuss 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 [5]. 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 modeling task. Third party annotation of great size can be done only in an automated way and it should be done according to an ontology.

Our annotation ontology grows out of URW3XG uncertainty ontology and extends some features needed for annotation. Below we show a part of our annotation ontology in Fig. 1. We start 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.

Web information extraction splits pages to dominantly tabular and/or textual. Uncertainty issues connected with information extraction (and annotation) from tabular pages were discussed in [1]. Extraction of textual pages will use techniques described in [6]. Both approaches (and any other approach) generate a level of (un)certainty they have about their annotations. Also users, human or agents, can review these uncertainties and provide feedback about them. 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 will be discussed in this presentation.

5 Pracovni seznam abstraktu

5.1 Dedek, Eckhardt, Galambos, Vojtas: Datakon 2008

[17]

Cílem je podat přehled vývoje zpracování informací na webu s důrazem na Sémantický web. Prvním klasifikačním hlediskem je, zda informace na webu jsou určeny pouze pro lidského konzumenta nebo také pro strojové zpracování (agent, služba). Vize Sémantického webu vidí ve strojovém zpracování webu možnost, jak obsah celého webu zpřístupnit všem. Základním stavebním kamenem Sémantického webu je RDF datový model s metadaty ve formě RDF schématu a OWL ontologií, vše standardizováno konsorciem W3C. Alternativní cesta Web 2.0 předkládá chytřejší uživatelská rozhraní a proprietární "mashup" webových dat. Nezanedbatelný je i podíl lidské práce (školené i neškolené) a automatizovaných procesů potřebný pro každou z těchto alternativ. Extrémní snaha vytvořit Sémantický web "od nuly" naráží na problém nejednotnosti ontologií, jejich různé kvality a potřeby jejich (automatizovaného) mapování. Na závěr představíme náš projekt sémantizace webu jako procesu, ve kterém klíčovou roli hraje extrakce dat z webu a anotace webových zdrojů.

6 Conclusion

We have presented a

6.1 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. In: Uncertainty Issues and Algorithms in Automating Process Connecting Web and User. Volume 5327 of Lecture Notes in Computer Science. Springer Verlag (2008)
2. 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
3. Dědek, J., Vojtáš, P.: Extrakce informací z textově orientovaných zdrojů webu. In Snášel, V., ed.: Znalosti 2008. (2008) 331–334
4. Dědek, J., Vojtáš, P.: Exploitation of linguistic tools in semantic extraction - a design. In Klopotek, M., Przepiórkowski, A., Wierzchoń, S., Trojanowski, K., eds.: Intelligent Information Systems XVI, Zakopane, Poland, Academic Publishing House EXIT (2008) 239–247
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., 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
7. 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
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. Eckhardt, A., Vojtáš, P.: Considering data-mining techniques in user preference learning. In: 2008 International Workshop on Web Information Retrieval Support Systems. (2008)
16. 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)
17. Dědek, J., Eckhardt, A., Vojtáš, P., Galamboš, L.: Sémantický web. In: DATAKON 2008, Brno (2008)