Web Information Extraction Systems for Web Semantization

Jan Dedek

Department of software engineering, Faculty of Mathematics and Physics
Charles University in Prague, Czech Republic
dedek@ksi.mff.cuni.cz

Abstract. *In this paper we present a survey of web information extraction systems and semantic annotation platforms. The survey is concentrated on the problem of employment of these tools in the process of web semantization. We compare the approaches with our own solutions and propose some future directions in the development of the web semantization idea.*

1 Introduction

There exist many extraction tools that can process web pages and produce structured machine understandable data (or information) that corresponds with the content of a web page. This process is often called Web Information Extraction (WIE). In this paper we present a survey of web information extraction systems and we connect these systems with the problem of web semantization.

The paper is structured as follows. First we sketch the basic ideas of semantic web and web semantization. In the next two sections methods of web information extraction will presented. Then description of our solutions (work in progress) will continue. And finally just before the conclusion we will discuss the connection of WIE systems with the problem of web semantization.

1.1 The Semantic Web in Use

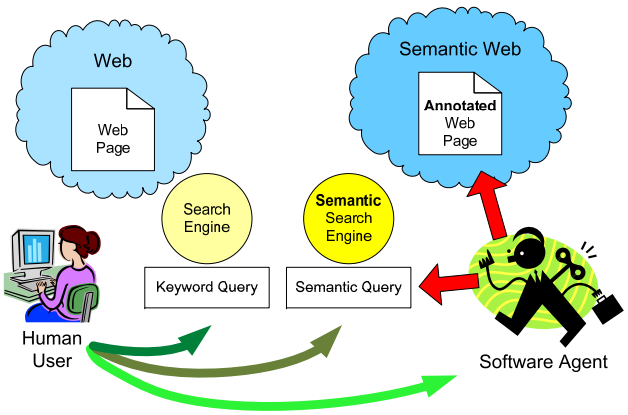

Fig. 1. The Semantic/Semantized Web in Use

The idea of the Semantic Web [4] (World Wide Web dedicated not only to human but also to machine – software agents) is very well known today. Let us just shortly demonstrate its use with respect to the idea of Web Semantization (see in next section).

The Fig. 1 shows a human user using the (Semantic) Web in three possible manners: a keyword query, a semantic query and by using a software agent. The difference between the first two manners (keyword and semantic query) can be illustrated with the question: "Give me a list of the names of E.U. heads of state." This example from interesting article [16] by Ian Horrocks shows the big difference between use of a semantic query language instead of keywords. In the semantic case you should be given exactly the list of names you were requesting without having to pore through results of (probably more then one) keyword queries. Of course the user have to know the syntax of the semantic query language or have a special GUI[1] at hand.

The last and the most important possibility (in the semantic or semantized setting) is to use some (personalized) software agent that is specialized to tasks of some kind like planning a business trip or finding the most optimal choice from all the relevant job offers, flats for rent, cars for sale, etc.

Both the semantic querying and software agents engagement is actually impossible to realize without any kind of adaptation of the web of today in the semantic direction.

1.2 Web Semantization

The idea of Web Semantization [9] consist in gradual enrichment of the current web content as an automated process of third party annotation for making at least a part of today's web more suitable for machine processing and hence enabling it intelligent tools for searching and recommending things on the web (see [3]).

The most strait forward idea is to fill a semantic repository with some information that is automatically extracted from the web and make it available to

[1] Such handy GUI can be found for example in the KIM project [20].

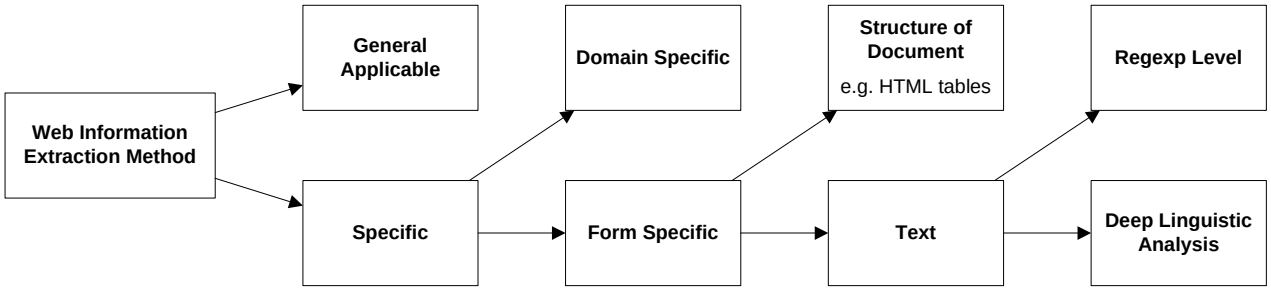

Fig. 2. Division of extraction methods

software agents so they could access to the web of today in semantic manner (e.g. through semantic search engine).

The idea of a semantic repository and a public service providing semantic annotations was experimentally realized in the very recognized work of IBM Almaden Research Center: the SemTag [13]. This work demonstrated that an automated semantic annotation can be applied in a large scale. In their experiment they annotated about 264 million web pages and generated about 434 millions of semantic tags. They also provided the annotations as a *Semantic Label Bureau* – a HTTP server providing annotations for web documents of 3rd parties.

2 Web Information Extraction

The task of a web information extraction system is to transform the web pages into program-friendly structures such as a relational database. There exists a rich variety of Web Information Extraction systems. The results generated by distinct tools usually can not be directly compared since the addressed extraction tasks are different. The extraction tasks can be distinguished according several dimensions: the task domain, the automation degree, the techniques used, etc. These dimensions are analyzed in detail in the recent publications [6] and [18]. Here we will concentrate on a little bit more specific division of WIE according to the needs of the Web Semantization (see in Sect. 5). The division is demonstrated on the Fig. 2 and should not be considered as disjoint division of the methods but rather as emphasization of different aspects of the methods. For example many extraction methods are domain and form specific at the same time.

The distinguishing between general applicable methods and the others that have meaningful application only in some specific setting (specific domain, specific form of input) is very important for Web Semantization because when we try to produce annotations in large scale, we have to control which web resource is suitable for which processing method (see in Sect. 5).

2.1 General Applicable

The most significant (and probably the only one) generally applicable IE task is so called *Instance Resolution Task*. The task can be described as follows: Given a general ontology, find all the instances from the ontology that are present in the processed resource. This task is usually realized in two steps: (1) Named Entity Recognition (see in Sect. 3.1), (2) Disambiguation of ontology instances that can be connected with the found named entities. Success of the method can be strongly improved with coreference resolution (see in Sect. 3.1).

Let us mention several good representatives of this approach: the SemTag application [13], the KIM project [20] and the PANKOW annotation method [7] based on smart formulation of Google API queries.

2.2 Domain Specific

Domain and from specific IE approaches are the typical cases. More specific information is more precise, more complex and so more useful and interesting. But the extraction method has to be trained to each new domain separately. This usually mens indispensable effort.

A good example of domain specific information extraction system is SOBA [5]. This complex system is capable to integrate different IE approaches and extract information from heterogeneous data resources, including plain text, tables and image captions but the whole system is concentrated on the single domain of football. Next similarly complex system is ArtEquAKT [1], which is entirely concentrated on the domain of art.

2.3 Form Specific

Beyond general applicable extraction methods there exist many methods that exploit specific form of the input resource. The linguistic approaches usually process text consisting of natural language sentences. The structure-oriented approaches can be strictly oriented

on tables [19] or exploit repetitions of structural patterns on the web page [21] (such algorithm can be only applicable to pages that contain more than one data record), and there are also approaches that use the structure of whole site (e.g. site of single web shop with summary pages with products conected with links to pages with details about single product) [17].

3 Information Extraction from Text-based Resources

In this section we will discuss the information extraction from textual resources.

3.1 Tasks of Information Extraction

There are classical tasks of text preprocessing and linguistic analysis like

Text Extraction – e.g from HTML, PDF or DOC,
Tokenization – detection of words, spaces, punctuations, etc.,
Segmentation – sentence and paragraph detection,
POS Tagging – part of speech assignment, often including lemmatization and morphological analysis,
Syntactic Analysis (often called linguistic *parsing*) – assignment of the grammatical structure to given sentence with respect to given linguistic formalism (e.g. formal grammar),
Coreference Resolution (or *anaphora resolution*) – resolving what a pronoun, or a noun phrase refers to. These references often cross boundaries of a single sentence.

Besides these classical general applicable tasks, there are further well defined tasks, which are more closely related to the information extraction. These tasks are domain dependent. These tasks were widely developed in the MUC-6 conference 1995 [15] and considered as semantic evaluation in the first place. These information extraction tasks are:

Named Entity Recognition: This task recognizes and classifies named entities such as persons, locations, date or time expression, or measuring units. More complex patterns may also be recognized as structured entities such as addresses.
Template Element Construction: Populates templates describing entities with extracted roles (or attributes) about one single entity. This task is often performed stepwise sentence by sentence, which results in a huge set of partially filled templates.
Template Relation Construction: As each template describes information about one single entity, this tasks identifies semantic relations between entities.
Template Unification: Merges multiple elementary templates that are filled with information about identical entities.
Scenario Template Production: Fits the results of Template Element Construction and Template Relation Construction into templates describing pre-specified event scenarios (pre-specified "queries on the extracted data").

Appelt and Israel [2] wrote an excellent tutorial summarizing these traditional IE tasks and systems built on them.

3.2 Information Extraction Benchmarks

Contrary to the WIE methods based on the web page structure, where we (the authors) do not know about any well established benchmark for these methods[2], the situation in the domain of text based IE is fairly different. There are several conferences and events concentrated on the support of automatic machine processing and understanding of human language in text form. Different research topics as text (or information) retrieval[3], text summarization[4] are involved.

On the filed of information extraction, we have to mention the long tradition of the Message Understanding Conference[5] [15] starting in 1987. In 1999 the event of *Automatic Content Extraction (ACE) Evaluation*[6] started, which is becoming a track in the Text Analysis Conference (TAC)[7] this year (in 2009).

All these events prepare several specialized datasets together with information extraction tasks and play an important role as information extraction benchmarks.

4 Our Solutions

4.1 Extraction Based on Structural Similarity

Our first approach for the web information extraction is to use the structural similarity in web pages containing large number of table cells and for each cell a link to detailed pages. This is often presented in web shops and on pages that presents more than one object

[2] It is probably at least partially caused by the vital development of the presentation techniques on the web that is still well in progress.
[3] e.g. Text REtrieval Conference (TREC) `http://trec.nist.gov/`
[4] e.g. Document Understanding Conferences `http://duc.nist.gov/`
[5] Briefly summarized in `http://en.wikipedia.org/wiki/Message_Understanding_Conference`.
[6] `http://www.itl.nist.gov/iad/mig/tests/ace/`
[7] `http://www.nist.gov/tac`

(product offer). Each object is presented in a similar way and this fact can be exploited.

As web pages of web shops are intended for human usage creators have to make their comprehension easier. Acquaintance with several years of web shops has converged to a more or less similar design fashion. There are often cumulative pages with many products in a form of a table with cells containing a brief description and a link to a page with details about each particular product.

Our main idea is to use a DOM tree representation of the summary web page and by breadth first search encounter similar subtrees. The similarity of these subtrees is used to determine the data region – a place where all the objects are stored. It is represented as a node in the DOM tree, underneath it there are the similar sub-trees, which are called data records.

We[8] have developed and implemented this idea [14] on the top of Mozilla Firefox API and experimentally tested on table pages from several domains (cars, notebooks, hotels). Similarity between subtrees was Levenshtein editing distance (for a subtree considered as a linear string), learning thresholds for decision were trained.

4.2 Linguistic Information Extraction

Our second approach [11, 12, 10] for the web information extraction is based on deep linguistic analysis. We have developed a rule-based method for extraction of information from text-based web resources in Czech and now we are working on its adaptation to English. The extraction rules correspond to tree queries on linguistic (syntactic) trees made form particular sentences. We have experimented with several linguistic tools for Czech, namely Tools for machine annotation – PDT 2.0 and the Czech WordNet.

Our present system captures text of web-pages, annotates it linguistically by PDT tools, extracts data and stores the data in an ontology. We have made 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.

To avoid the need of manual design of extraction rules we focused on the data extraction phase and made some promising experiments [8] with the machine learning procedure of Inductive Logic Programming for automated learning of the extraction rules.

This solution is directed to extraction of information which is closely connected with the meaning of text or meaning of a sentence.

[8] Thanks go mainly to Dušan Maruščák and Peter Vojtáš.

5 The Web Semantization Setting

In this section we will discuss possibilities and obstructions connected with the employment of web information extraction systems in the process of web semantization.

One aspect of the realization of the web semantization idea is the problem of integration of all the components and technologies starting with web crawling, going through numerous complex analyses (document preprocessing, document classification, different extraction procedures), output data integration and indexing, and finally implementation of query and presentation interface. This elaborate task is neither easy nor simple but today it is solved in all the extensive projects and systems mentioned above.

The novelty that web semantization brings into account is the cross domain aspect. If we do not want to stay with just general ontologies and general applicable extraction methods then we need a methodology how to deal with different domains. The system has to support extension to a new domain in generic way. So we need a methodology and software to support this action. This can for example mean: to add a new ontology for the new domain, to select and train proper extractors and classifiers for the suitable input pages.

5.1 User Initiative and Effort

An interesting point is the question: Whose effort will be used in the process of supporting new domain in the web semantization process? How skilled such user has to be? There are two possibilities (demonstrated on the Fig 3). The easier one is that we have to employ very experienced expert who will decide about the new domain and who will also realize the support needed for the new domain. In the Fig 3 this situation is labeled as *Provider Initiated* and *Provider Trained* because the expert works on the side of the system that provides the semantics.

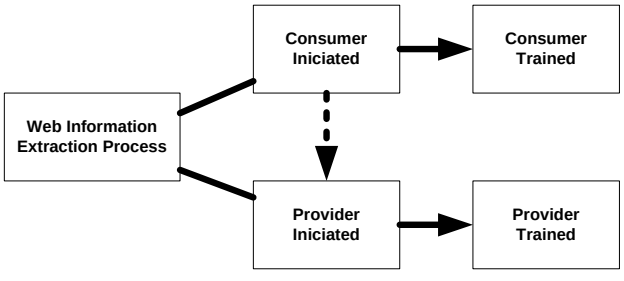

Fig. 3. User initiative and effort

The second possibility is to enable ordinary users form outside to add a new domain to the semanti-

zation system. Such user is probably interested in semantic data form the domain so we call such user *Consumer*.

A cooperation of a consumer (possibly a domain expert) and a provider (system expert) on the support of the new domain can be considered as hybrid approach. This is represented with the dashed arrow in the Fig 3.

6 Conclusion and Future Work

In this paper we tried to show the complexity of the problem of web semantization in connection with the possibilities of web information extraction systems. Future work goes in several directions:

- Future development of WIE tools and work on their adaptability to new domains.
- Integration of WIE tools to the web semantization system.
- Development of the methodology and software to support the extension of the semantization system to a new domain for a non-expert user.

Acknowledgments

This work was partially supported by Czech projects 1ET100300517, 201/09/H057 Czech Science Foundation and MSM-0021620838.

References

1. Harith Alani, Sanghee Kim, David E. Millard, Mark J. Weal, Wendy Hall, Paul H. Lewis, and Nigel R. Shadbolt. Automatic ontology-based knowledge extraction from web documents. *IEEE Intelligent Systems*, 18:14–21, 2003.
2. Douglas E. Appelt and David J. Israel. Introduction to information extraction technology. A tutorial prepared for IJCAI-99, Stockholm, Schweden, 1999.
3. Tim Berners-Lee. The web of things. *ERCIM News - Special: The Future Web*, 72:3, January 2008.
4. Tim Berners-Lee, James Hendler, and Ora Lassila. The semantic web, a new form of web content that is meaningful to computers will unleash a revolution of new possibilities. *Scientific American*, 284(5):34–43, May 2001.
5. Paul Buitelaar, Philipp Cimiano, Anette Frank, Matthias Hartung, and Stefania Racioppa. Ontology-based information extraction and integration from heterogeneous data sources. *Int. J. Hum.-Comput. Stud.*, 66(11):759–788, 2008.
6. Chia-Hui Chang, M. Kayed, M. R. Girgis, and K. F. Shaalan. A survey of web information extraction systems. *Knowledge and Data Engineering, IEEE Transactions on*, 18(10):1411–1428, 2006.
7. Philipp Cimiano, Siegfried Handschuh, and Steffen Staab. Towards the self-annotating web. In *WWW '04: Proceedings of the 13th international conference on World Wide Web*, pages 462–471, New York, NY, USA, 2004. ACM.
8. Jan Dědek, Alan Eckhardt, and Peter Vojtáš. Experiments with czech linguistic data and ILP. In Filip Železný and Nada Lavrač, editors, *ILP 2008 - Inductive Logic Programming (Late Breaking Papers)*, pages 20–25, Prague, Czech Republic, 2008. Action M.
9. Jan Dědek, Alan Eckhardt, Peter Vojtáš, and Leo Galamboš. Sémantický web. In Václav Řepa and Oleg Svatoš, editors, *DATAKON 2008*, pages 12–30, Brno, 2008.
10. Jan Dědek and Peter Vojtáš. Computing aggregations from linguistic web resources: a case study in czech republic sector/traffic accidents. In Cosmin Dini, editor, *Second International Conference on Advanced Engineering Computing and Applications in Sciences*, pages 7–12. IEEE Computer Society, 2008.
11. Jan Dědek and Peter Vojtáš. Exploitation of linguistic tools in semantic extraction - a design. In Mieczyslaw Klopotek, Adam Przepiórkowski, Slawomir Wierzchoń, and Krzysztof Trojanowski, editors, *Intelligent Information Systems XVI*, pages 239–247, Zakopane, Poland, 2008. Academic Publishing House EXIT.
12. Jan Dědek and Peter Vojtáš. Linguistic extraction for semantic annotation. In Costin Badica, Giuseppe Mangioni, Vincenza Carchiolo, and Dumitru Burdescu, editors, *2nd International Symposium on Intelligent Distributed Computing*, volume 162 of *Studies in Computational Intelligence*, pages 85–94, Catania, Italy, 2008. Springer-Verlag.
13. Stephen Dill, Nadav Eiron, David Gibson, Daniel Gruhl, R. Guha, Anant Jhingran, Tapas Kanungo, Sridhar Rajagopalan, Andrew Tomkins, John A. Tomlin, and Jason Y. Zien. Semtag and seeker: bootstrapping the semantic web via automated semantic annotation. In *WWW '03: Proceedings of the 12th international conference on World Wide Web*, pages 178–186, New York, NY, USA, 2003. ACM.
14. Alan Eckhardt, T. Horváth, D. Maruščák, R. Novotný, and Peter Vojtáš. *Uncertainty Issues and Algorithms in Automating Process Connecting Web and User*, volume 5327 of *Lecture Notes in Computer Science*. Springer Verlag, 2008.
15. Ralph Grishman and Beth Sundheim. Message understanding conference-6: a brief history. In *Proceedings of the 16th conference on Computational linguistics*, pages 466–471, Morristown, NJ, USA, 1996. Association for Computational Linguistics.
16. Ian Horrocks. Ontologies and the semantic web. *Commun. ACM*, 51(12):58–67, 2008.
17. Kristina Lerman, Lise Getoor, Steven Minton, and Craig Knoblock. Using the structure of web sites for automatic segmentation of tables. In *SIGMOD '04: Proceedings of the 2004 ACM SIGMOD international conference on Management of data*, pages 119–130, New York, NY, USA, 2004. ACM.
18. Bing Liu. *Web Data Mining*. Springer-Verlag, 2007.

19. David Pinto, Andrew Mccallum, Xing Wei, and Bruce W. Croft. Table extraction using conditional random fields. In *SIGIR '03: Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval*, pages 235–242, New York, NY, USA, 2003. ACM Press.
20. Borislav Popov, Atanas Kiryakov, Damyan Ognyanoff, Dimitar Manov, and Angel Kirilov. Kim – a semantic platform for information extraction and retrieval. *Nat. Lang. Eng.*, 10(3-4):375–392, 2004.
21. Hongkun Zhao, Weiyi Meng, Zonghuan Wu, Vijay Raghavan, and Clement Yu. Fully automatic wrapper generation for search engines. In *WWW Conference*, pages 66–75, 2005.