Information Extraction from Web Pages Based on Their Visual Representation

Ruslan R. Fayzrakhmanov*

Database and Artificial Intelligence Group
Institute of Information Systems, TU Vienna
Favoritenstrasse 9, A-1040 Vienna, Austria
fayzrakh@dbai.tuwien.ac.at

Abstract. This research is dedicated to enhancing the efficiency of web information extraction and web accessibility. The motivation behind the research, its aim and objectives are presented, and the performed work on developing web page model for information extraction is described. We also present work on making extracted information accessible to blind users, providing them with the means to navigate and access required information quickly. We also present our ongoing research on creating efficient methods and approaches for information extraction from the proposed model. There are two main approaches considered: 1) development of the library which provides required functionality to the programmer; 2) development of declarative Datalog-like language for information extraction.

Keywords: web information extraction, web page, wrapper, web accessibility.

1 Introduction

The Web is an enormous repository of information. It plays an important role in business, politics, science, and our everyday life. Web pages are the main components of the Web, presenting information in semi-structured and unstructured forms, using well-known standards, such as HTML and XHTML. These forms of representation and CSS are solely used for specifying visual formatting, and they are convenient forms for storing and transferring information through the Internet. But HTML, XHTML as well as DOM tree are not designed to present semantics and data types on a web page. As is generally known, most of the contemporary information extraction systems consider only source code or DOM tree, which, besides the characteristics mentioned, are exposed to frequent change. The semantics of a web page are hidden in its visual representation, where — beside textual and multimedia contents — colour, size, style (of text), and the relative position of elements play and important role. Regardless of the

* Supported by the Erasmus Mundus External Cooperation Window Programme of the European Union.

A. Harth and N. Koch (Eds.): ICWE 2011 Workshops, LNCS 7059, pp. 342–346, 2011.

template used for web page generation and its source code, humans are able to distinguish and identify different web objects on different web pages (e.g. news article, navigation menu) that allow us to put forward a hypothesis about existence of some permanent characteristics in web objects such as relative position and relative size.

The *idea* behind this research, which is a part of ongoing ABBA[1] porject, is based on utilizing positional information, spatial expansion of visual objects, and their visual characteristics for information extraction. In addition, in the research we solve the problem of accessibility of the extracted information for blind users. The *aim* of the research is the development of methods and tools to enhance the efficiency of information extraction for web pages based on their visual representation, and to present it in a form accessible for blind users. To achieve this goal, the following *objectives* were formulated:

1. Review and analysis of related work (100% completed).
2. Development of a web page model based on its visual representation (90% completed).
3. Development of methods for information extraction from the proposed web page model (10% completed).
4. Development of methodology for navigation through the extracted information for blind users (90% completed).
5. Development of the information extraction system on the basis of proposed methods (10% completed).
6. Development of a navigation system according to proposed methodology (90% completed).
7. Analysis of efficiency of proposed methods of information extraction and navigation (10% completed).

This research is carried out under the supervision of Prof. Reinhard Pichler, Dr. Robert Baumgartner (TU Vienna).

2 A Web Page Model

For the tasks of information extraction and web page understanding within the scope of the ABBA project, a web page model was developed, describing its visual characteristics and taking into account its DOM tree [6]. Continuing the research, we propose a *web page model* as a conjunction of its *geometrical* (GM) and *logical* (LM) models (cf. Fig. 1).

A **GM** is an ontological model and is formed as a result of the analysis of web page visual representation (its CSS model), generated by the browser's layout engine. A GM represents visual information in a form convenient for both information extraction and web page understanding. The *geometric object* (GO) of the GM has a rectangular shape and wraps some part of the web page canvas.

[1] The ABBA project (Advanced Barrier-free Browser Accessibility) is sponsored by the Austrian Forschungsförderungsgesellschaft FFG under grant 819563.

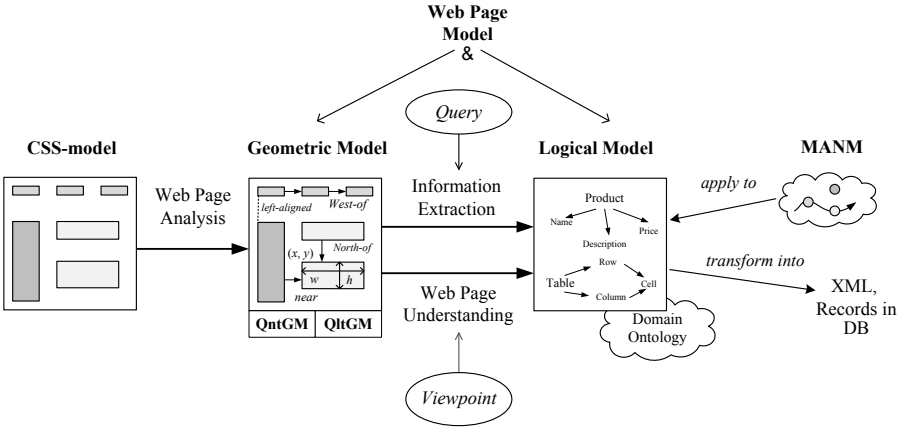

Fig. 1. Diagram represents the process of automatic creation of a web page model

It can correspond to some CSS box. The main attributes of a GO are features of a background (colour, background image), border (width, colour, style), contained text (colour, font size and style) as well as drawing order, which is used to define visibility of overlapped GOs. It can be calculated according to the painting order (W3C specification [1]).

Depending on the type of information representation, we define *quantitative* GM (QntGM) and *qualitative* GM (QltGM). Spatial relationships, such as distance and direction, are defined between GOs in the QntGM and expressed quantitatively (in pixels and angles respectively). In the QltGM, besides distance and direction, there are alignment and topological relationships (we use RCC8 [5]) expressed qualitatively via linguistic variables. These relationships are widely used both for graphical user interfaces [10] and positional information representation in GIS [4].

A **LM** is an ontological model and is formed as a result of the process of web page understanding or information extraction (cf. Fig. 1). In the first case, an LM represents a semantics of web objects on the web page with the required level of detail. In the second case, an LM describes the necessary part of the web page according to the request. An LM set a correspondence between GOs of the GM and concepts of the applied domain ontology. Thus, this solution contributes also to the Semantic Web development, providing us with necessary semantic metadata annotations [9].

An LM can be transformed to the XML format or stored in a database, but in this research we focus on accessibility of extracted information for blind users.

3 Development of Information Extraction Methods

Information extraction from the GM is represented as a gradual process of successive refinement of extracted information characteristics and its extraction [8].

For instance, to extract posts in a web forum, we need to indicate the location (in our case it is center) and occupied area of the object to be extracted. A post can be further described as a rectangular area which contains a textual message occupying a major part, and also contains an icon at the top-left corner, etc. When extracting items from the navigation menu, for instance, we first define its approximate area of occurrence and define the menu as a horizontally or vertically oriented list of textual elements. To specify extracted web object or its parts, one can use positional information, an HTML type of corresponding element in the source code, and its CSS style, provided by the GM of the web page.

We consider two solutions for information extraction. The **first** one involves developing the Java library, which provides necessary functionality for a programmer to create a wrapper. Algorithms implemented in the library should be efficient, giving a posibility to the programmer to utilize all potential of GM. The **second** solution involves developing declarative Datalog-like extraction language. Its predicates will not be evaluated over an extensional database of the facts representing the GM, but directly over the GM, represented as an ontological model. It will make the performance more efficient. This solution is similar to the Lixto solution, where ELog language is used and which predicates are evaluated over DOM tree [3]. This language is very efficient and useful for automatic wrapper generation according to the performed specifications of extracted information by the user, using only GUI. Thus, it does not require any programming skills from the end user.

4 Navigating Extracted Information

Within the scope of the ABBA project in which the author participate, multiaxial navigation model (MANM) along with the methodology of navigation were developed to make web pages more accessible to a blind user [7], [2]. In this research, the MANM is used for making extracted information (concepts in the LM) accessible.

The main component of the MANM is the axis (cf. Fig 2), which is a sequence of web page model elements to be read. The MANM is provided both with the possibility to navigate on the axis (e.g., news titles) and change axis. Moving from one to another axis can be performed by its selection from the

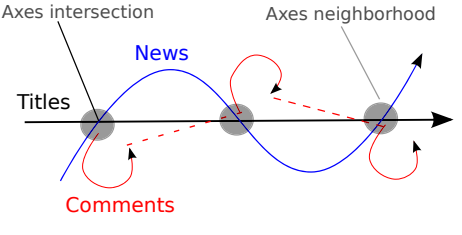

Fig. 2. Example of the quantitative geometrical model

set of all available axes, from the set of axes intersecting current element, from the spatial, or semantic neighbourhood of current element. For this reason, we consider semantic relationship between elements defined in the LM and their spatial relations defined in the GM.

5 Conclusion

This paper describes the current state of Ruslan R. Fayzrakhmanov's research. The work is dedicated to problems in information extraction from web page visual representation and web accessibility. The model of web pages for information extraction and the multi-axial navigation model are presented. Further work on the development of methods of information extraction is also described.

References

1. Cascading Style Sheets Level 2 Revision 1 (CSS 2.1) Specification (2009), http://www.w3.org/TR/2009/CR-CSS2-20090908/
2. Baumgartner, R., Fayzrakhmanov, R.R., Holzinger, W., Krüpl, B., Göbel, M.C., Klein, D., Gattringer, R.: Web 2.0 vision for the blind. In: Proc. of Web Science Conference 2010 (WebSci 2010), Raleigh, USA, p. 8 (2010)
3. Baumgartner, R., Flesca, S., Gottlob, G.: The Elog Web Extraction Language. In: Nieuwenhuis, R., Voronkov, A. (eds.) LPAR 2001. LNCS (LNAI), vol. 2250, pp. 548–560. Springer, Heidelberg (2001)
4. Clementini, E., Di Felice, P., Hernández, D.: Qualitative representation of positional information. Artificial Intelligence 95(2), 317–356 (1997)
5. Cohn, A.G.: Qualitative spatial representation and reasoning techniques, pp. 1–30. Springer, Berlin (1997)
6. Fayzrakhmanov, R.R., Göbel, M.C., Holzinger, W., Krüpl, B., Baumgartner, R.: A Unified ontology-based web page model for improving accessibility. In: Proc. WWW 2010, pp. 1087–1088. ACM, New York (2010)
7. Fayzrakhmanov, R.R., Göbel, M.C., Holzinger, W., Krüpl, B., Mager, A., Baumgartner, R.: Modelling Web navigation with the user in mind. In: Proc. W4A 2010, Raleigh, USA, p. 4 (2010)
8. Gottlob, G., Koch, C., Baumgartner, R., Herzog, M., Flesca, S.: The Lixto data extraction project - back and forth between theory and practice. In: Transformation, pp. 1–12. ACM, Paris (2004)
9. Kashyap, V., Bussler, C., Moran, M.: The Semantic Web. Semantics for Data and Services on the Web. Springer, Berlin (2008)
10. Kong, J., Zhang, K., Zeng, X.: Spatial graph grammars for graphical user interfaces. ACM Transactions on Computer-Human Interaction 13(2), 268–307 (2006)