Journal of Visual Languages and Computing 21 (2010) 23–32

Contents lists available at ScienceDirect

Journal of Visual Languages and Computing

journal homepage: www.elsevier.com/locate/jvlc

Visual extraction of information from web pages

Giuseppe Della Penna *, Daniele Magazzeni, Sergio Orefice

Department of Computer Science, University of L'Aquila, Via Vetoio, I-67100 Coppito, L'Aquila, Italy

ARTICLE INFO

Article history:
Received 24 June 2008
Received in revised form
8 June 2009
Accepted 22 June 2009

Keywords:
Information extraction
Web visual information search
Spatial relations
Query languages
User interfaces for search interaction

ABSTRACT

In this paper we present a graphical software system that provides an automatic support to the extraction of information from web pages. The underlying extraction technique exploits the *visual appearance* of the information in the document, and is driven by the *spatial relations* occurring among the elements in the page. However, the usual information extraction modalities based on the web page structure can be used in our framework, too. The technique has been integrated within the Spatial Relation Query (SRQ) tool. The tool is provided with a graphical front-end which allows one to define and manage a library of spatial relations, and to use a SQL-like language for composing queries driven by these relations and by further semantic and graphical attributes.

1. Introduction

The explosive growth of popularity of the World Wide Web has resulted in a huge amount of information sources on the Internet. Therefore, the data extraction from the web is becoming an important topic and poses many challenges in devising effective methodologies to search, access, and integrate information.

The process of knowledge acquisition from the Web can be divided conceptually into three consecutive steps [1]: information retrieval (IR) which aims to locate documents containing data and information relevant to a certain query; information extraction (IE) which aims to extract from the located documents relevant information that appear in certain semantic or syntactic relationships; information integration (II) which aims to compose the individual pieces of information extracted from the documents and to build an integrated view (see Fig. 1).

In particular, IE tries to process the relevant information found on the documents in order to make it available to structured queries. Most often, information extraction systems are customized for specific application domains, and require manual or semi-automatic training sessions. A survey of some existing IE tools can be found in Section 2. These tools are commonly based on approaches relying on structural information (e.g., HTML code), training examples or pattern matching.

In this paper we propose an approach based on the *visual appearance* of the information, conceived as its *rendering* through a web browser. This allows one to shift the IE problem from the low level of code (HTML, CSS, etc.) to the higher level of visual features, providing a paradigm of the kind "what you see drives your search" that supports a natural query formulation.

In particular, in our approach the extraction of information is driven by the spatial arrangement of the elements in the web page (e.g., *spatial relations* such as "right of" or "included in"). The syntax of these spatial relations is based on the formal framework of *visual language classes* defined in [2]. Here, each class characterizes a family of visual languages based upon the nature of their graphical objects and composition rules. In particular, in the present paper we use the *box syntactical*

* Corresponding author.
E-mail addresses: giuseppe.dellapenna@univaq.it (G. Della Penna), daniele.magazzeni@univaq.it (D. Magazzeni), sergio.orefice@univaq.it (S. Orefice).

1045-926X/$ - see front matter © 2009 Elsevier Ltd. All rights reserved.
doi:10.1016/j.jvlc.2009.06.001

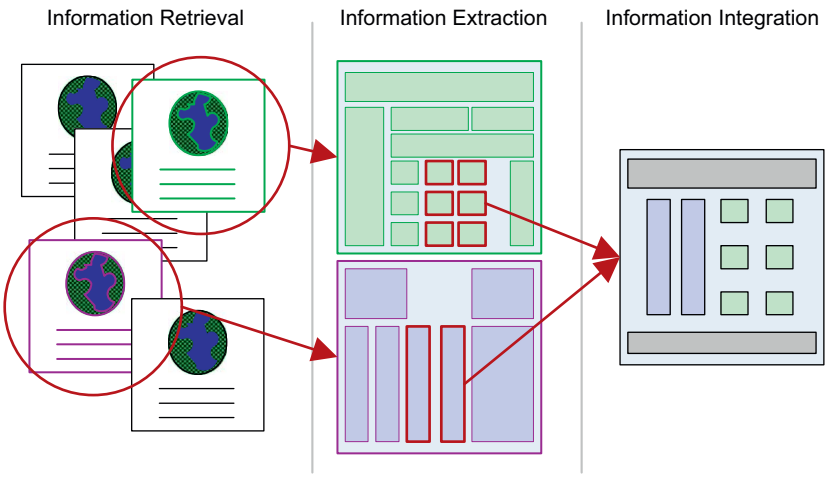

Fig. 1. Web information acquisition phases.

model which represents the more general class of geometric-based visual languages.

To this aim, we have designed a SQL-like query language, namely SRQL (Spatial Relation Query Language), which allows one to write queries based on the visual arrangement of the web page contents in an intuitive way. Moreover, the SRQL can also use further semantic and graphical attributes of the web page elements to refine the queries, thus providing a framework where the information extraction can be performed integrating spatial relations, visual attributes, textual content and document structure.

The box spatial relation formalism and the SRQL have been implemented within the SRQ (Spatial Relation Query) software tool. The tool is provided with a graphical front-end which allows one to define and manage a library of spatial relations, compose and execute queries, and export the extracted information for further analysis and manipulation. For instance, it is possible to extract the tables on the right of an image, or the links between two text blocks, or the text of a certain color within a paragraph beginning with a determined word, or the text below the first/last image, etc.

In particular, the SRQ tool is suitable to extract information from web pages having the following characteristics, which are often hard to handle by the common IE approaches:

- Highly variable or very complex HTML structure.
- Client-side dynamically modified content, especially in the case of Javascript-powered pages (e.g., using AJAX techniques).
- Pages whose source code is not easily accessible.

The paper is organized as follows. Related work is summarized in Section 2 together with some comparisons between our approach and other techniques and tools. In Section 3 we give some basic notions on the formalism of spatial relations underlying the tool. The architecture of the SRQ tool is illustrated in Section 4, where we also describe the SRQL syntax. Section 5 contains a case study and some considerations on the tool experimentation. To conclude, some final remarks are outlined in Section 6.

2. Related work

In the last years much research has been done for supporting information extraction on the web, and several approaches have been proposed. In particular, recent surveys of web data extraction tools [3–5] have singled out the following categories:

(a) *Wrapper induction tools* (e.g., DEPTA [6], IEPAD [7] WIEN [8], LIXTO Visual Wrapper [9], SoftMealy [10], STALKER [11]) that are based on the structural information and formatting features of the web pages, and generate extraction rules derived from a given set of training examples or pattern discovery techniques.
(b) *HTML-aware tools* (e.g., LIXTO [12], W4F [13], XWrap [14,15], RoadRunner [16]) that rely on the structural information of the web pages, too, and use HTML parse trees for creating extraction rules.
(c) *Modeling-based tools* (e.g., DEByE [17], Robosuite [18], NoDoSE [19]) that locate in the web pages portions of data conforming to a predefined structure provided according to a set of modeling primitive as tuples or lists.
(d) *NLP-based tools* (e.g., WHISK [20], RAPIER [21], SRV [22]) that work on phrases and sentences elements within the web pages to derive extraction rules, by applying techniques such as filtering, part-of-speech tagging and lexical semantic tagging.
(e) *Ontology-based tools* (e.g., WeDaX [23], BYU [24], On-To-Knowledge [25]), that rely on the data and not on the structure of the source documents. These tools use ontology to locate constants in the page and to construct objects with them.

Moreover, recent research efforts are focusing on the development of:

(f) *Specific-targeted tools*, e.g., WiNTs [26], that automatically produces wrappers that can be used to extract