Web Information Extraction Using Eupeptic Data in Web Tables*

Wolfgang Gatterbauer, Bernhard Krüpl, Wolfgang Holzinger, and Marcus Herzog

Database and Artificial Intelligence Group
Institute of Information Systems, Vienna University of Technology
Favoritenstr. 9-11, 1040 Vienna, Austria
{gatter, kruepl, holzing, herzog}@dbai.tuwien.ac.at

Abstract. By leveraging on the redundant information on the Web, we are building a Web information extraction system that concentrates on eupeptic data in Web tables. We use the term eupeptic to describe such representations of information that allow for easy interpretation of the subject–predicate–object nature of individual data items. The system mimics a human approach to information gathering. It explicitly uses visual cues on rendered Web pages to locate tabular data; it uses keywords to identify relevant chunks of data that gets processed on a deeper level; and it expands its initial search to include more pages when it spots eupeptic data.

Keywords: Web information extraction, Web tables, Visual approach to information extraction, S–P–O triples.

1 Introduction

The Web contains a wealth of product information with detailed attribute descriptions. The goal of our work is to gather such entity information and transform it into a coherent format for further processing. We have chosen the digital camera domain as the first test environment for our system.

We perceive relevant information as instances of subject–predicate–object (S–P–O) triples similar to the data model of RDF describing resources in the form of resource–property–value statements. Such S–P–O triples appear in various syntactic expressions, either implicitly like in "Canon Ixus IIs can take pictures with a resolution up to 2048 × 1536 pixels" or explicitly like in "(Canon Ixus IIs–resolution–2048×1536)".

In our approach, we focus on eupeptic data, using the term *eupeptic* to describe such representations of information that allow for easy interpretation of

* This research is supported in part by the Austrian Federal Ministry for Transport, Innovation and Technology under the FIT-IT contract FFG 809261, and the DOC Scholarship Program of the Austrian Academy of Sciences.

Vojtěch Svátek, Václav Snášel (Eds.): RAWS 2005, pp. 41–48, ISBN 80–248–0864–1.

the S–P–O nature of individual data items. In particular tables containing reasonable short pieces of text exhibit such "eupeptic qualities".

By limiting ourselves to the use of eupeptic data, we are exploiting the redundancy of the Web. Our goal is not to find and extract all appearances of relevant statements from our domain on the Web. Instead, we strive to extract as much *unique* information as possible, for which we introduce the measure *unique recall* (r_{u}). Whereas the traditional measure *recall* (r) assesses the ability of a system to extract all appearances of relevant information from a certain domain, unique recall describes the ability to extract all relevant information stripped of all redundancy.

Defining the term redundancy (ρ) as the number of appearances of a certain piece of information on the Web, the two measures are related according to $r_{\mathrm{u}} = 1 - (1 - r)^{\rho}$ under the simplified assumption of equal redundancy among all pieces of information. For domains with many instances of the same piece of information on the Web, e.g. the digital camera domain, r_{u} becomes high even with low r, e.g. $r_{\mathrm{u}} \cong 0.9$ with $\rho = 20$ and $r = 0.1$.

With this idea in mind, we mimic a human approach by selecting Web pages that "please" the system for the subsequent analysis. Hence the two logical stages of Web information extraction, information retrieval (IR) and information extraction (IE), are not detached modules in our approach, but connected with a feedback loop (see Fig. 1).

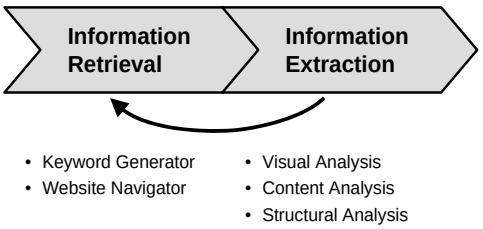

Fig. 1. Web information extraction as an interconnected process between IR and IE.

We will outline each of the two main stages in the following two sections, followed by an overview of related work and our conclusions.

2 Information Retrieval

The information retrieval stage of our data harvesting process uses domain knowledge and page layout understanding for finding and selecting relevant Web pages containing eupeptic data. It operates in three steps: finding relevant Web pages by querying existing internet search engines, analyzing retrieved pages for their suitability for extraction and crawling identified Web sites to gather further pages.

Generating Queries for Internet Search Engines. In the first step, the system uses suitable keywords for entities and their attributes to generate query strings for search engines. We strive to submit the most descriptive and distinctive search terms to narrow the result set. The system assesses the quality of each word by its global frequency (measuring the hits from a search engine) as well as by judgement gathered from previous experience, learning which words represent important facets of entities or produce good results.

Choosing Web Pages. The pages found in the first step represent a sample of the information available on their respective Web sites. In the second step we decide if these samples contain data that our information extraction subsystem can process — by submitting these pages to the extraction component. Only Web sites that prove fit for extraction remain in the next step.

Crawling Web Sites. The aim of the third step is to crawl sites chosen by the the information extraction component to enrich the collection of relevant Web pages. Figure 2 shows the (simplified) graph spanned by the hyperlinks of a typical Web site in our testing domain. The valuable pages we seek are represented on the outer perimeter. Inside, with the site's root in the center, is a tightly interwoven maze of hyperlinks forming an almost complete sub-graph. It is therefore necessary to develop good heuristics guiding the crawling process to be able to reach the target pages in adequate time. At the moment we use the tree structure of the page URLs for hints but other schemes, such as analyzing certain graph properties, will also be considered.

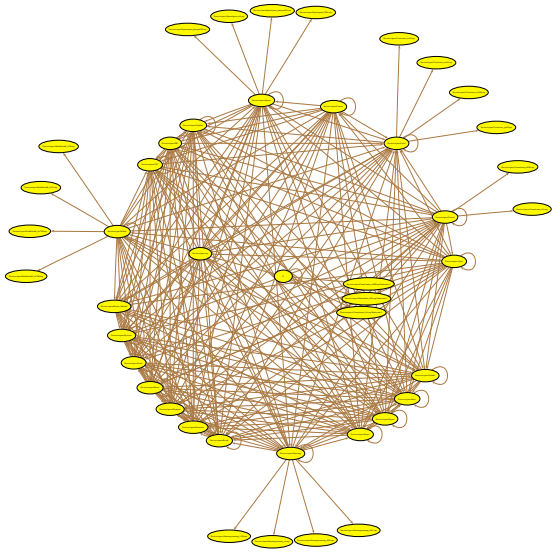

Fig. 2. Typical site graph; outward nodes represent pages containing eupeptic data.

3 Information Extraction

The goal of this stage is to test the retrieved Web pages for eupeptic qualities and extract S–P–O statements from those found to be suitable. Feedback is given to the IR stage, which crawls Web sites with identified eupeptic data for more relevant Web pages.

We use term S–P–O triples instead of the more common terminology of extracting attribute name–value pairs for given items to emphasize that we not only look for attribute information of certain items, but also for subjects that exhibit a certain predicate–object relationship.

Web pages distinguish themselves from purely text based documents by the fact that valuable information is encoded in the formatting of the documents themselves. Figure 1 implies that we explicitly consider the visual representation of data to be of equal importance to content and structure in the extraction process. Hence, our principal dimensions are *visual appearance*, *content*, and *structure*.

Although the interaction between these three dimensions is heavily interrelated and not simply linear, the overall extraction process can be considered as a successive analysis along these three dimensions (see Fig. 3). In the following subsections, we discuss these steps in more detail.

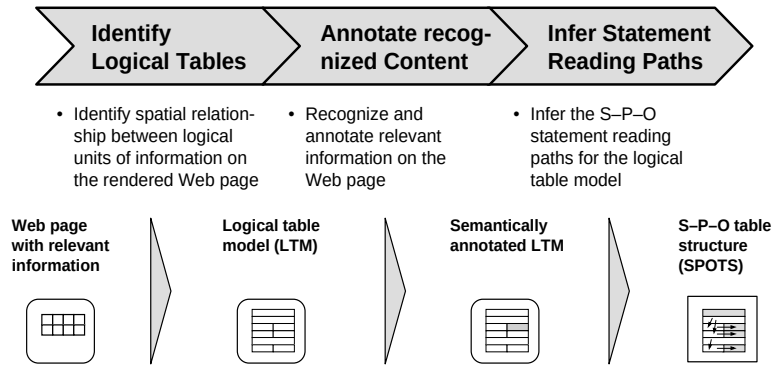

Fig. 3. Three principal steps of the information extraction stage.

3.1 Identifying Logical Tables

When human authors want to address human readers, they usually think in terms of text flow and layout. But formatting a document is not just a question of design. Layout can also provide important semantic information: related things are often formatted the same way, and semantically similar items are often grouped together. Tables, the strictest form of formatting, provide readers with an even clearer way to understand what the author wants to tell them by grouping the items they contain along two or more dimensions.

In Web documents, many different forms of source code can lead to the same visual rendering. With CSS becoming more and more popular, the interpretation of (X)HTML code becomes even more complicated, as can be judged from the difficulties many Web browsers have in adhering to the standards. Therefore, it makes sense to delegate the task of source code interpretation to a tool that is known to handle it well, the Mozilla Gecko renderer. By interfacing with Gecko, we derive the positional information of all text boxes on a Web page and analyze the page as perceived by a human reader instead of operating on the obfuscated (X)HTML and possibly CSS source code. By doing this, we work on a level of representation that is closer to the original author's meaning of the document (see Fig. 4).

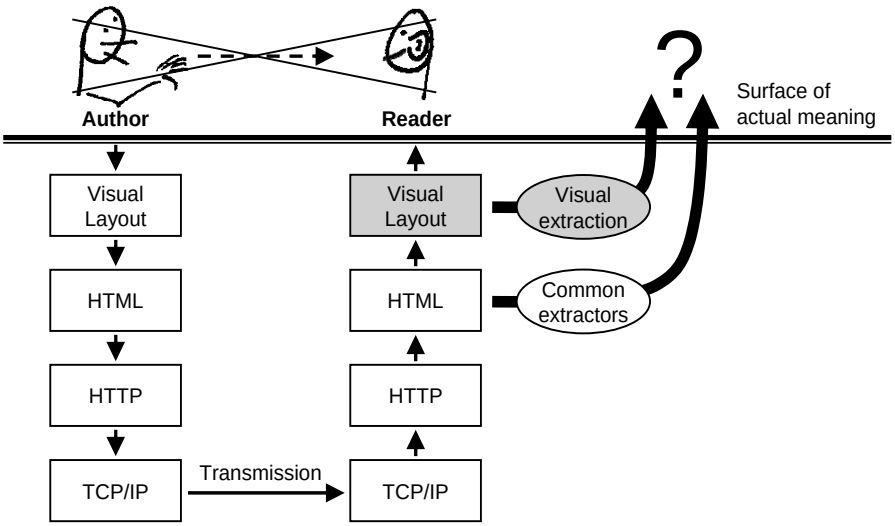

Fig. 4. The visual approach to information extraction: extracting from rendered Web pages instead of deeper levels of abstraction.

We segment every input page to locate tables. Since we are working on the visual layer, we apply a variant of an algorithm widely known in the OCR community, namely the X-Y cut algorithm [5]. This algorithm creates horizontal and vertical whitespace density graphs for the whole page and finds the most prominent gap in the two graphs (see Fig. 5). It uses this gap to cut the page into two segments. The algorithm then runs recursively on these two segments until a final threshold is reached. We enhance the segmentation process by providing information from the other component of the extraction stage, the content extraction component. The output of this step is the logical table model (LTM), an explicit model of the spatial relationship between the logical units of information as understood by the human reader.

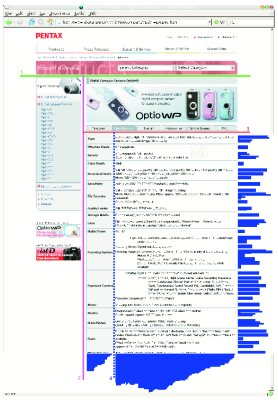

Fig. 5. Segmentation by cutting at white space density graph gaps.

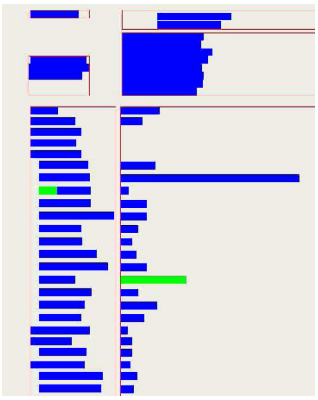

Fig. 6. NERs spot relevant concepts on a page (green).

3.2 Annotating Recognized Content

Information extraction is about getting facts out of documents. The concepts of a particular domain of interest have to be specified in advance. We formalize this specification of facts in a domain model that we are currently constructing. This model consists of a taxonomy of concepts known in the domain of interest and a set of sample instances of this model. Named Entity Recognizers (NERs) use this information to locate entities on the page under investigation (see Fig. 6).

Again, this is in line with the human approach that we try to follow. When looking for relevant information, a human reader not only uses visual cues to segment a page into logical pieces; she also scans the page quickly to identify terms of particular interested. Only after she has spotted one of these words does he start to read and analyze on a deeper grammatical level.

Our decision to concentrate on Web pages with tabular data has another important consequence: data rich, eupeptic tables usually do not contain full sentences, but only collections of words or short text fragments. Therefore, no sophisticated NLP understanding of sentence structures is necessary when extracting data from such tables. Instead, we examine what one could term "micro-grammars" of table cells, for example options for enumerations (e.g. using different separation characters) or specification of number ranges (in particular regular expressions).

3.3 Inferring Statement Reading Paths

The last step of the principal extraction process is the inference of the S–P–O statement reading paths, by which we refer to the implied routes through cells that are semantically linked and collectively express the information we are looking for. Figure 7 shows the three inputs for what we refer to as the SPOTS

(S–P–O table structure) recognition: the logical table model from the table identification, cell metadata from the Mozilla Gecko renderer and sample identified cells from the content annotation. This step is followed by a straightforward transformation into a predefined XML schema that serves as the interface for a subsequent information integration stage with a deeper semantic analysis of the individual components.

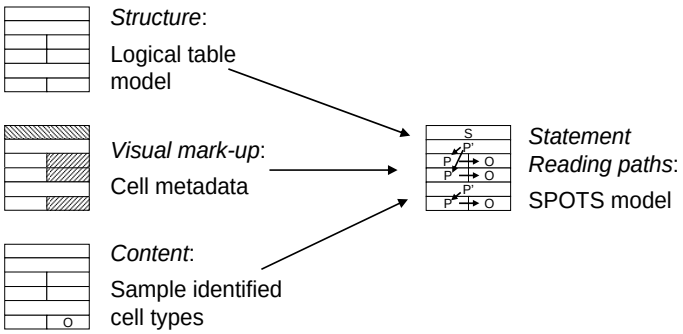

Fig. 7. Input and output of the S–P–O table structure recognition.

Predicates often appear in an implied hierarchy with the hierarchy depending upon the domain ontology chosen by the author of the respective source. As a result, the actual S–P–O statements often do not appear in the form of triples, but rather in quadruples of logical units. The flattening of the hierarchy will be taken care of in the information integration stage. The information extraction stage considers the statements in the form of generalized S–P–O statements such as S–(P,P')–O (see Fig. 7).

Such generalized S–P–O statements appear in Web tables in a limited number of syntactical patterns. We are currently analyzing example SPOTS models from our domain for recurring micro patterns along the three dimensions of the input. Because of its assumed robustness, our goal is rather a bottom-up agent based table grammar inference then a top-down approach that would directly map LTMs to SPOTs.

At the time of writing this paper, this last step has not yet been concretized in its full extent. Simplified tests of subtasks do seem very promising, yet details of implementation, smooth integration of the components and thorough tests of this approach remain to be done.

4 Related Work

Current table literature often refers to the PhD thesis of Hurst [1] and its comprehending table model of five layers. Pivk et al. [2] use a slightly adapted table model with four layers and present the semantics of a table in F-Logic frames instead of a relational database. Both publications mention the graphical

dimension of tables, but explicitly disregard this dimension for their table understanding mechanisms. Neither uses the metadata information of Web table cells for inferring the reading order. Costa e Silva et al. [3] describe how the introduction of feedback loops in the information extraction process can help improve the table recognition process. We use such a feedback loop between the IR and IE stages and between the three dimensions of our extraction process. The X-Y cut algorithm used for visual segmentation has been originally devised in the OCR community [4] and has been applied to Web pages by our group [5]. Embley et al. use ontologies for to generate wrappers for Web pages [6]. We also build on domain knowledge for extraction, but employ concept spotters to identify eupeptic parts on web pages with data aligned in S–P–O triples. The idea of extracting S–P–O triples was also mentioned by Svatek et al. in [7], where the authors propose searching for certain expressions of visual patterns of the P–O relationship in the HTML source code.

5 Conclusion

In this paper, we have presented a rough outline of our work in progress, a Web information extraction method that focuses on eupeptic data in Web tables. Extracting from eupeptic data follows a human approach, a paradigm that underlies all of our work: in exploiting the redundancy on the Web, we concentrate on those pages where S–P–O triples of information can be identified and extracted by applying rather simple table grammars. Moreover, the composition of query string keywords, the use of domain knowledge for quick spotting of relevant text chunks and the extraction on top of the rendered view of Web pages follow this line of thought.

Acknowledgement. We would like to thank the anonymous reviewers for their fruitful comments that helped us improve the clarity of the paper.

References

1. M. Hurst: The interpretation of Tables in Texts. *PhD thesis, University of Edinburgh*, 2000.
2. A. Pivk, P. Cimiano, Y. Sure: From Tables to Frames. *preprint*, 2005.
3. A. Costa e Silva, A. Jorge, L. Torgo: Design of an end-to-end method to extract information from tables. 2003, www.niaad.liacc.up.pt/ amjorge/docs/tablesmm.pdf
4. G. Nagy and S. Seth: Hierarchical representation of optically scanned documents. *Proc. of the 7th Int. Conf. on Pattern Recognition*, pp. 347–349, 1984.
5. B. Krüpl, M. Herzog, W. Gatterbauer: Using Visual Cues for Extraction of Tabular Data from Arbitrary HTML Documents. Poster track, *WWW 2005*, Chiba 2005.
6. D. W. Embley: Toward Tomorrow's Semantic Web – An Approach Based on Information Extraction Ontologies. *Position Paper for Dagstuhl Seminar*, 2005.
7. V. Svatek, J. Braza, V. Sklenak: Towards Triple-Based Information Extraction from Visualy-Structured HTML Pages. Poster Track, *WWW 2003*, Budapest 2003.