Robust Web Content Extraction

Marek Kowalkiewicz
Department of Management Information Systems
The Poznan University of Economics
Al. Niepodleglosci 10
60-967 Poznan, Poland
+48 (61) 8543631
marek@kowalkiewicz.net

Maria E. Orlowska
Department of Information Technology and Electrical Engineering
The University of Queensland
Brisbane, St. Lucia
QLD 4072, Australia
+61 (7) 33652989
maria@itee.uq.edu.au

Tomasz Kaczmarek
Department of Management Information Systems
The Poznan University of Economics
Al. Niepodleglosci 10
60-967 Poznan, Poland
+48 (61) 8543631
tomek@kie.ae.poznan.pl

Witold Abramowicz
Department of Management Information Systems
The Poznan University of Economics
Al. Niepodleglosci 10
60-967 Poznan, Poland
+48 (61) 8543381
witold@abramowicz.pl

ABSTRACT
We present an empirical evaluation and comparison of two content extraction methods in HTML: absolute XPath expressions and relative XPath expressions. We argue that the relative XPath expressions, although not widely used, should be used in preference to absolute XPath expressions in extracting content from human-created Web documents. Evaluation of robustness covers four thousand queries executed on several hundred webpages. We show that in referencing parts of real world dynamic HTML documents, relative XPath expressions are on average significantly more robust than absolute XPath ones.

Categories and Subject Descriptors
H.5.4 [**Information Interfaces and Presentation**]: Hypertext/Hypermedia – *architectures, navigation, theory.*

General Terms
Measurement, Performance, Reliability, Experimentation

Keywords
Content Extraction, Robustness, Wrappers, Evaluation.

1. INTRODUCTION
Web content extraction technologies, such as one presented in this paper, may be applied to personalize Web content. Our work was motivated by the fact that dynamic webpages often contain semantically irrelevant content blocks. We envisioned a tool that was able to extract only relevant blocks from webpages. An important concern of such tool was robustness of content extraction, or in other words, ability to retrieve desired content even if source documents' structure changes or if the content itself is changed.

For content extraction we applied XPath queries over XHTML documents. XPath language allows for absolute and relative queries. The contribution of the paper is in comparison of the robustness of the two methods. Although intuition suggests an answer, no such extensive comparative study has been undertaken so far. We present results of this comparison based on our previous work (content extraction and aggregation system - [2]).

WWW 2006, May 23–26, 2006, Edinburgh, Scotland.
ACM 1-59593-323-9/06/0005.

There is a number of technologies to extract information from webpages. For a state of the art analysis, see previously mentioned paper [2] and an extensive study by Laender et al. [3]. Robustness of content extraction methods, with a focus on XPath, has been presented in a work by Abe and Hori [1]. The authors proposed four content anchoring approaches and tested them on documents from four Web sites. The samples used in theory study were too small to be statistically valid. However the authors make a significant step towards establishing an understanding of the applicability of content extraction methods in real world Web content extraction. The study did not give a definite answer on which XPath method is the most robust Web content extraction.

2. QUERYING THE WEB WITH XPATH
We define web content as text or images visible in the browser, separated into visually or semantically distinguishable sections. This visual and semantic separation is obtained through proper embedding of the content within HTML elements. HTML is represented as a labeled ordered rooted tree (accessible as DOM tree). Nodes of this tree can be accessed via XPath expressions, which are therefore suitable for content extraction. XPath expressions are constructed by giving element names and their sibling order on subsequent tree levels. XPath expressions can be divided into two categories: absolute and relative ones, in regards to the node that starts the path. Absolute addressing expressions start with "/", denoting document root node; evaluation of relative addressing expressions, on the other hand, starts at a node that can be arbitrarily selected (anchor node) (Figure 1).

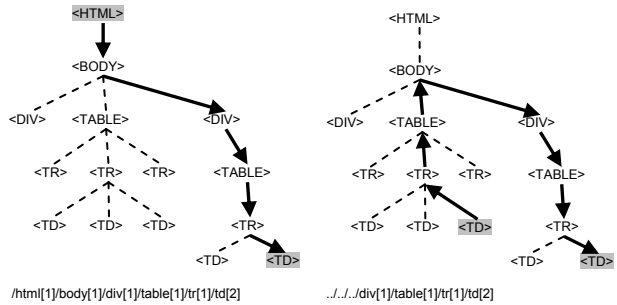

Figure 1 Absolute (left) and relative (right) path expressions with root/reference and extracted nodes highlighted

Relative addressing, applied to content extraction, requires a reliable method to indicate candidate anchor node. We believe relative addressing is more robust, since absolute addressing can be sensitive to changes in the overall document structure (like introduction of

ads, changing placement of sections etc.) as well as to changes in details of document structure (redesign within a section). Properly used, relative addressing is vulnerable only to the second type of changes. Additionally, relative addressing more accurately imitates human behavior. Users locate relevant information on the webpage by seeking visual clues like headers or table captions. The way the information is presented on the Web (similar to paper publications), makes significant changes in the relative positioning of information and its descriptive elements unlikely. Relative addressing requires introducing a reference element - a textual element that is a starting point for evaluation of the relative XPath expression.

We empirically examined the question of any significant difference in robustness of the two methods. We addressed the question by performing extensive empirical evaluations.

3. EXPERIMENT AND RESULTS

We adopted the following procedure for measuring robustness. First, we defined a set of content sources (randomly selected from Google Directory's branches of news and general purpose portals). We used archived pages (instances) of these sources for the whole of 2004. We asked three persons, not involved in the creation of the system, to create 5 absolute and 5 relative queries for each of the sources. The queries were always defined for the first instance of a given content source in 2004, and the users were shown only the mentioned instance of a website. We tested the queries on all instances of each content source. We defined robustness for a single user query (absolute and relative) as an extraction success ratio measured as the percentage of correctly extracted content blocks throughout the whole year. The robustness of a method (absolute, relative) for a given content source was defined as an average robustness of corresponding queries. The overall method robustness was measured as a weighted average of robustness measures for each source. We weighted the measures by considering the number of archived instances for each website.

To compare robustness, we ran and evaluated almost four thousand queries, grouped into 70 groups (each group queried one content block over several instances). Our experiments show that the relative XPath expressions are on average more robust than the absolute XPath expressions. Absolute queries were robust in 47%, while relative enjoyed 76% robustness. Furthermore, our study shows, that if website layout does not change dramatically (we defined "dramatic change" intuitively, as a complete change of site's layout), the relative XPath method's robustness is on average more than 95% (while absolute XPath stays below 75%) – these results were obtained by restricting the set of tested websites. Surprisingly, our experiments show, that the length of XPath expression (number of nodes) is not significantly correlated with its robustness. It is however worth mentioning that the vulnerability of both methods should be attributed to different factors. Change in document structure was always the reason for failure of absolute queries. Relative queries failed due to inability to locate anchoring element or (supposedly rarely) due to changes in detailed document layout (structure changes close to the leaves of DOM tree). Moreover we found out that there were many cases where absolute queries failed at some point and did not recover for subsequent instances of a given content source, probably due to major webpage redesign. At the same time relative queries were able to recover after just a few instances. We think this phenomenon may be explained by changes that occur in the textual content of the reference element or slight adjustments in document layout, that are more likely to occur on the websites that are edited by humans.

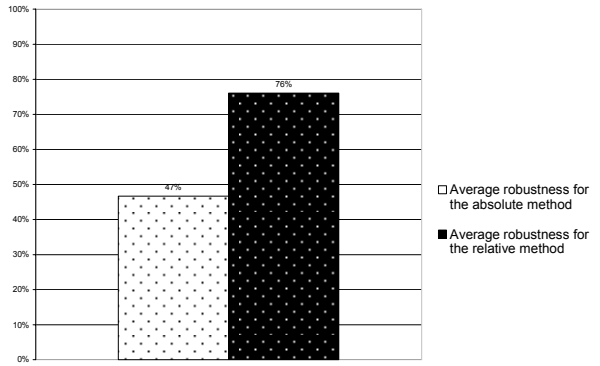

Figure 2 Overall comparison of robustness of the two evaluated XPath methods

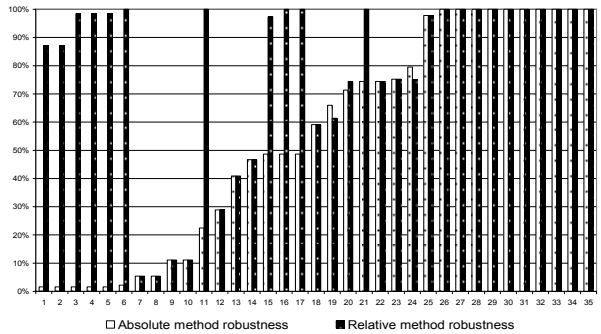

Figure 3 Comparison of absolute and relative query robustness for 70 query groups, sorted by absolute robustness. Relative expressions are at least as robust as absolute ones in all but two cases

4. CONCLUSIONS

Our research question was whether there is a significant difference in robustness of the extraction methods based on absolute and relative XPath expressions. With our proof-of-concept implementation and empirical study conducted on statistically significant number of webpages we proved that relative XPath expressions are significantly more robust.

5. REFERENCES

[1] Abe, M. and Hori, M. Robust Pointing by XPath Language: Authoring Support and Empirical Evaluation. in *Proceedings of 2003 Symposium on Applications and the Internet (SAINT 2003), 27-31 January 2003*, IEEE Computer Society, Orlando, FL, USA, 2003, 156-165.

[2] Kowalkiewicz, M., Orlowska, M., Kaczmarek, T. and Abramowicz, W. Towards more personalized Web: Extraction and integration of dynamic content from the Web. in *Proceedings of the 8th Asia Pacific Web Conference APWeb 2006*, Harbin, China, 2006.

[3] Laender, A.H.F., Ribeiro-Neto, B.A., Silva, A.S.d. and Teixeira, J.S. A brief survey of web data extraction tools. *ACM SIGMOD Record*, *31* (2). 84-93.

Research supported by a Polish state research grant 1 H02D 084 28 (Marek Kowalkiewicz) and a Marie Curie Transfer of Knowledge Fellowship MTKD-CT-2004-509766 (Maria Orlowska).