

Expected Utility of Content Blocks in Web Content Extraction

Marek Kowalkiewicz
Department of Management Information Systems
The Poznan University of Economics
M.Kowalkiewicz@kie.ae.poznan.pl

Abstract

In this paper we discuss the possible application of new concepts in web content extraction: utility assessment, utility annealing, and dynamic aggregated document generation. After analysis of the state of the art in web content extraction, results of a survey study among Polish managers are presented. The discussion covers a web content extraction system with possible extensions that may help tackle the information overload problem. The discussed extensions go beyond current state of the art. Utility assessment considers economical view on value of information, while utility annealing allows for removing content blocks that cover information already acquired from other content blocks. Due to the existing content block extraction technology and new concepts proposed in the paper, it is possible to dynamically generate aggregated documents.

1. Introduction

Currently we are facing an overburdening growth of the number of reliable information sources on the Internet. The quantity of information available to everyone via Internet is dramatically growing each year [1]. At the same time, temporal and cognitive resources of human users are not changing, therefore causing a phenomenon of information overload.

World Wide Web is one of the main sources of information for decision makers (reference to my research). However our studies show that, at least in Poland, the decision makers see some important problems when turning to Internet as a source of decision information. One of the most common obstacles raised is distribution of relevant information among many sources, and therefore need to visit different Web sources in order to collect all important content and analyze it.

A few research groups have recently turned to the problem of information extraction from the Web [2]. The most effort so far has been directed toward collecting data from dispersed databases accessible via web pages (related to as data extraction or information extraction from the Web) and towards

understanding natural language texts by means of fact, entity, and association recognition (related to as information extraction). Data extraction efforts show some interesting results, however proper integration of web databases is still beyond us. Information extraction field has been recently very successful in retrieving information from natural language texts, however it is still lacking abilities to understand more complex information, requiring use of common sense knowledge, discourse analysis and disambiguation techniques.

1.1. Vision

Since automated information extraction do not fulfill expectations, especially when analyzing largely unstructured business documents, we believe that an interesting approach towards reducing the phenomenon of information overload is to provide methods and tools for content extraction and aggregation. One such method and tool has been proposed by Kowalkiewicz, Orłowska, Kaczmarek and Abramowicz [3].

So far the tools and methods of content extraction and aggregation do not consider an important fact. Namely the information needs are dynamically changing, and facing two information items of the same expected relevance, the relevance of one item may dramatically fall as soon as a user views the other (and acquires requested information) [4].

A problem of content aggregation methods is that the aggregation may use a limited space for presenting aggregated views. Our vision is to introduce a new concept of content utility to Web content extraction and aggregation systems that could be used as an extension of traditional relevance approach in order to present users content blocks of some significance. Such an approach would lead to optimal utilization of browser display areas.

1.2. Research Challenges

The research challenge of the work is to construct a strategy of assessing content block utility, ideally using already know methods originating from Information Retrieval and Economics fields.

1.3. Contribution

In this paper we show how a content utility assessment could improve users' experience while fulfilling their information needs. We also draw a preliminary vision of the method. Since the paper shows in-progress work, the discussion here should be treated as an invitation to commenting the work and possibly extending the concepts.

2. Web Content Extraction

Web content extraction is an interesting field, attracting many groups of researchers. Research is done as an answer to user needs, and its results are implemented in content extraction applications. In this section we analyze state of the art content extraction technologies, show results of a survey conducted among Polish managers, and describe a proof-of-concept application, myPortal, used to perform web content extraction experiments.

2.1. State of the art

Content extraction is understood as extracting complex, semantically and visually distinguishable information, such as paragraphs or whole articles from the Web. It borrows from information extraction methods used in the World Wide Web environment, and especially from Web data extraction methods. The most comprehensible survey of Web data extraction tools has been provided by Laender et al. [2], there are however other ones, also relevant to our study.

The WebViews system [5] is a GUI system that allows users to record a sequence of navigation and point interesting data in order to build a wrapper. User is able to point interesting data; however it is not clear how the query to document's data is generated. The system is limited to extracting data from tables. IEPAD [6] is a system used to automatically extract repetitive subsequences of pages (such as search results). It is interesting in the context of wrapper generation and content extraction. IEPAD uses PAT trees to identify repetitive substructures and is prone to specific types of changes in subsequent substructures (for instance changing attributes of HTML tags, additional symbols between tags). Annotea [7], on the other hand, is a system designed not for content extraction, but for its annotation. The work provides a description of an approach of addressing specific parts of HTML documents. The authors present the method on XML documents, implicitly assuming that the conversion from HTML to XML representation has been done. As the authors point themselves, the method is very sensitive to changes in the document, which makes it usable only in addressing content of static documents. eShopMonitor [8] is a complex system providing tools for monitoring content of Web sites. It consists of three components: crawling system, which retrieves interesting webpages; miner, allowing users to point interesting data and then extracting the data; and reporting systems, which executes queries on extracted data and then provides user with consolidated results. The miner uses XPath expressions to represent interesting data. ANDES (A Nifty Data Extraction Systems) [9] extracts structured data using XHTML and XSLT technologies. The author of this system decomposes the extraction problem into five sub-problems: website navigation, data extraction, hyperlink synthesis, structure synthesis, data mapping, and data integration. WysiWyg Web Wrapper Factory (W4F) [10] is a set of tools for automatic wrapper generation. It provides tools for generating retrieval rules and a declarative language for building extraction rules. W4F uses a proprietary

language, making it hard to integrate with other extraction systems. WebL [11] is a data extraction language. It is possible to represent complex queries (such as recursive paths and regular expressions) with it, however the language provides very limited means to address XML documents, particularly it doesn't support XSLT templates and XPath expressions. Chen, Ma, and Zhang [12] propose a system that clips and positions webpages in order to display them properly on small form factor devices. They use heuristics in order to identify potentially interesting content. Their clipping methods, according to a set of 10'000 analyzed HTML pages, behaves perfectly (no problems in page analysis and splitting) in around 55% of documents. Out of remaining 45%, some 35% percent documents cause problems in page splitting, and the final 10% generates errors in both page analysis and splitting.

Other possibly interesting systems include: WIDL [13], Ariadne [14], Garlic [15], TSIMMIS [16], XWRAP [17], and Informia [18]. It is important to note, that none of the mentioned systems was designed explicitly to extract previously defined content from dynamically changing webpages.

We have not found information on any system that would consider content utility in content extraction and aggregation. However, some approaches have been made. The most notable is a work of Anderson, Domingos, and Weld [19].

2.2. User needs specification

We have conducted a survey among Polish managers. The surveyed group included over 120 managers, undertaking Master of Business Administration studies. We asked the users several questions regarding their habits and remarks related to Internet use at workplace. The surveys were preceded by a short introduction, including overall information on web content extraction. The results of the survey, among others, showed that the Internet is an important source of information used at work (Figure 1).

Our research also showed, that the main concerns related to using Internet as an information source are connected with scattered nature of information (requiring visiting many sources in order to reach one information source) and overload of irrelevant content (implying usefulness of content extraction in order to weed out unneeded document parts). Figure 2 shows distribution of answers among the group of surveyed managers.

The survey included other questions, regarding frequency and duration of Internet sessions at work, number of regularly visited information sources, types of information sought in the Internet and a question about a potential impact of content aggregation on work efficiency of the surveyed (Figure 3).

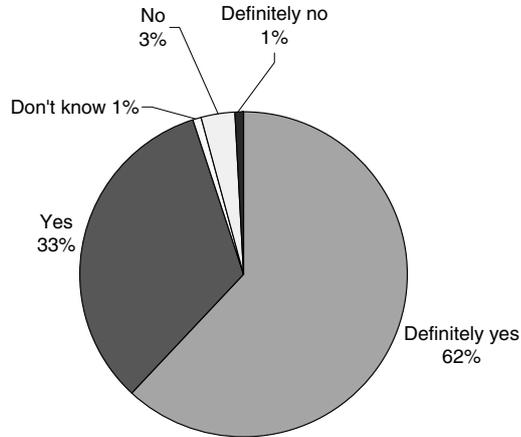


Figure 1. *Is the Web an important source of Information you use at work?*
Results of the survey distributed among Polish managers.

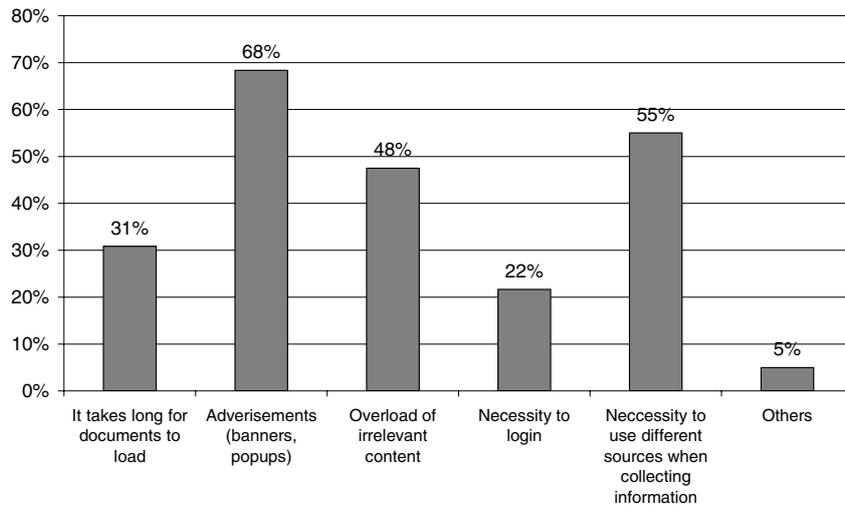


Figure 2. *Indicate the most important problems that impede the usage of Internet as an information source (please select up to three answers).* Results of the survey distributed among Polish managers.

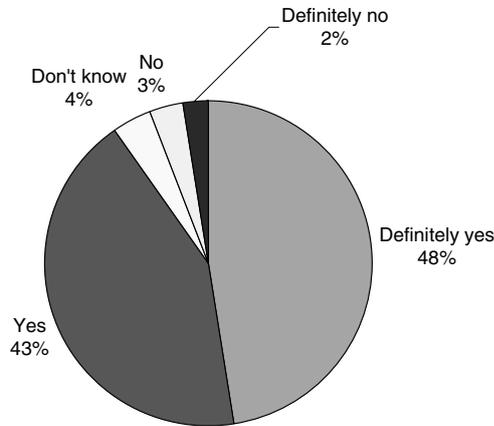


Figure 3. *Do you think that getting a consolidated report instead of a need to search for every piece of information on the Internet would improve your efficiency?* Results of the survey distributed among Polish managers.

The survey results indicate that the Internet is already a very important source of information for managers. The amount of time they spend while collecting information is a significant cost for organizations. Therefore research towards reducing time spent on information foraging is well justified.

2.3. myPortal

While researching how to improve the information collecting phase, we have developed a method of content block extraction from dynamic webpages, more robust than others available. The method has been implemented in a proof-of-concept application – myPortal [3]. MyPortal is a content extraction and aggregation system that provides a point-and-click interface allowing users to specify their information needs and thus build a tailored portal (dubbed myPortal) containing only previously chosen content blocks, able to obtain most recent content from the Web. The approach used in myPortal provides a significant increase in robustness, allowing users to create portals that present required information even when changes to structure occur.

myPortal utilizes XPath language, relies on relative paths, and provides capabilities of locating anchor points (not using XPath constructs). Such approach makes more robust pointing possible. We have tested myPortal robustness on several hundred webpages, using several thousand queries, and compared myPortal's method robustness to absolute XPath robustness. The robustness gain is over 60% [3]

The application constructs a web portal, where extracted content blocks are located (aggregated). The content blocks are currently located in selected regions. Since there may be many content block definitions in the system, and screen area is very limited, it would be desirable if the system was able to select, design and present a portal view with most important information. Portal redesign could be done as soon as any of the content blocks would be read

3. Content Block Utility

3.1. Economical aspects of information utility

Akerlof [20] was one of the first to put a stress on value of information in economical decision making. After that others – from both economics and information science – followed with important research [21]. An interesting approach to confronting economical models and the research field of information retrieval was presented by Varian [4]. He argued that the value of information can be measured: *the value of information is the increment in expected utility resulting from the improved choice made possible by better information* [4]. What is the most important conclusion for the IR community is that it is only new information that matters. Therefore acquisition of information from one document reduces relevance of another one, returned as a result of some user query. Therefore, post retrieval clustering of results may be a good approach towards evolution of IR systems by reducing the cognitive load and disambiguation. There is a number of research studies conducted in that area.

3.2. Content block utility

When considering content block extraction, one has to be aware of two facts. (1) In typical content extraction application, users are interested in different topics – their information needs are complex and most often include more than one query. (2) A typical web document consists of multiple content blocks (semantically distinguishable units), which again may be of different interest to users. A content extraction and aggregation system may extract content blocks from different web documents and place them in one view (Figure 4). Some of the content blocks may be of a higher importance to user. Some other blocks may duplicate the information.

An interesting research goal is to be able to assess the utility of content blocks. We are currently conducting studies on how to assess the “probability of relevance” in order to place content blocks in an aggregated document according to their relevance. We believe that it is possible to create a dynamic document that will change according to users’ behavior (and therefore changing utility of individual content blocks).

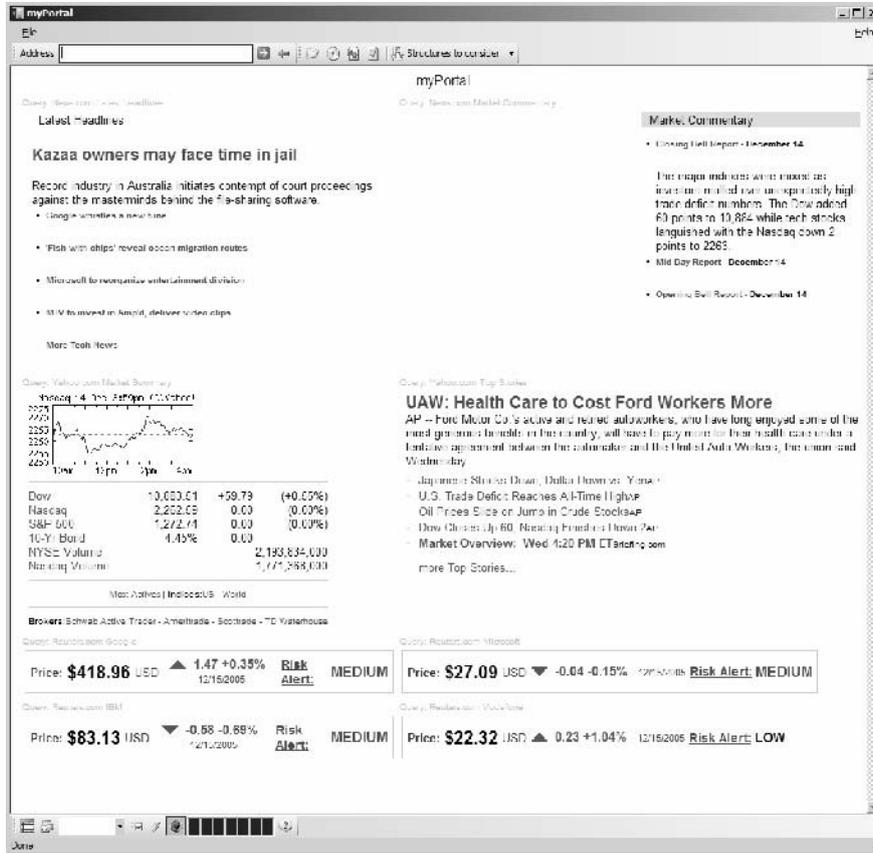


Figure 4. A sample screen from a content extraction and aggregation system myPortal aggregating News.com, Yahoo.com and Reuters.com content blocks [3].

3.3. Utility annealing

As Varian puts it [4], as soon as a user acquires information from one source, other ones, covering the same topic, become less relevant. When referring web content blocks, we call this phenomenon utility annealing. Whenever a user reads a content blocks, the aggregated document should be dynamically rearranged and most relevant content blocks should be moved to the top, while similar topics should be degraded (as their utility decreases). Such behavior requires using traditional IR methods such as assessing relevance and clustering results.

3.4. Utility driven content block selection and ordering

When assessing content block utility becomes feasible, a new application of content extraction and aggregation tools emerges. It is possible that far more content blocks are extracted than can be perceived by a user in a given time. A new document, aggregating all interesting content blocks may be constructed from scratch, and provided that a limit of a document size (considering users' cognitive limitations) is given – a new document with aggregated content can be created. Only content blocks that are above a certain utility threshold will be selected and those will be ordered according to their utility and then presented to users.

4. Conclusions

Current web content extraction systems are not flexible enough. Even though they aim at reducing the information overload problem, there is still a lot to be done in the field. One of the approaches that may be useful in web content extraction is to construct documents aggregating different web content blocks according to their utility. No work has been done so far in the topic. We believe that after suggesting a Web content extraction method along with a prototype system that proved to be more robust than the other ones used currently, it is possible to construct a method for dynamic content aggregation based on its utility. In this paper, after analyzing state of the art in content extraction, presenting results of our preliminary studies, we presented basic ideas underlying the concept. Further work includes development of utility assessment method and a technique for constructing dynamic aggregated documents with utility annealing. If successful, the practical implementation will be of use wherever large quantities of unstructured or semi-structured information are analyzed.

5. References

1. P. Lyman, H. R. Varian, K. Swearingen, P. Charles, N. Good, L. L. Jordan, and J. Pal, "How Much Information 2003?," School of Information Management and Systems, the University of California at Berkeley 2003.
2. A. H. F. Laender, B. A. Ribeiro-Neto, A. S. d. Silva, and J. S. Teixeira, "A brief survey of web data extraction tools," *ACM SIGMOD Record*, vol. 31, pp. 84-93, 2002.
3. M. Kowalkiewicz, M. Orłowska, T. Kaczmarek, and W. Abramowicz, "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.
4. H. R. Varian, "Economics and Search," in *SIGIR 1999*. Berkeley, California: ACM Press, 1999, pp. 1-5.

5. J. Freire, B. Kumar, and D. Lieuwen, "WebViews: Accessing Personalized Web Content and Services," in Proceedings of the 10th international conference on World Wide Web, V. Y. Shen, N. Saito, M. R. Lyu, and M. E. Zurko, Eds. Hong Kong: ACM Press New York, 2001, pp. 576-586.
6. C.-H. Chang and S.-C. Lui, "IEPAD: Information Extraction based on Pattern Discovery," in Proceedings of the 10th international conference on World Wide Web, V. Y. Shen, N. Saito, M. R. Lyu, and M. E. Zurko, Eds. Hong-Kong: ACM Press New York, 2001, pp. 681-688.
7. J. Kahan, M.-R. Koivunen, E. Prud'Hommeaux, and R. R. Swick, "Annotea: An Open RDF Infrastructure for Shared Web Annotations," in Proceedings of the 10th international conference on World Wide Web, V. Y. Shen, N. Saito, M. R. Lyu, and M. E. Zurko, Eds. Hong-Kong: ACM Press New York, 2001, pp. 623-632.
8. N. Agrawal, R. Ananthanarayanan, R. Gupta, S. Joshi, R. Krishnapuram, and S. Negi, "The eShopmonitor: A comprehensive data extraction tool for monitoring Web sites," IBM Journal of Research and Development, vol. 48, pp. 679-692, 2004.
9. J. Myllymaki, "Effective Web Data Extraction with Standard XML Technologies," in Proceedings of the 10th international conference on World Wide Web, V. Y. Shen, N. Saito, M. R. Lyu, and M. E. Zurko, Eds. New York, NY, USA: ACM Press, 2001, pp. 689-696.
10. A. Sahuguet and F. Azavant, "WysiWyg Web Wrapper Factory (W4F)," in Proceedings of the 8th International World Wide Web Conference, A. Mendelzon, Ed. Toronto: Elsevier Science, 2000.
11. T. Kistler and H. Marais, "WebL - A Programming Language for the Web," in Proceedings of the 7th International World Wide Web Conference. Brisbane, Australia, 1998.
12. Y. Chen, W.-Y. Ma, and H.-J. Zhang, "Detecting Web Page Structure for Adaptive Viewing on Small Form Factor Devices," in Proceedings of the 12th International World Wide Web Conference. Budapest, Hungary: ACM Press New York, 2003, pp. 225-233.
13. C. Allen, "WIDL: Application Integration with XML," World Wide Web Journal, vol. 2, 1997.
14. C. A. Knoblock, S. Minton, J. L. Ambite, N. Ashish, P. J. Modi, I. Muslea, A. G. Philpot, and S. Tejada, "Modeling Web Sources for Information Integration," in Proc. Fifteenth National Conference on Artificial Intelligence, 1998.
15. M. T. Roth and P. Schwarz, "Don't Scrap It, Wrap It! A Wrapper Architecture for Legacy Data Sources," in Proceedings of the 23rd VLDB Conference. Athens, Greece, 1997, pp. 266-275.
16. J. Ullman, S. Chawathe, H. Garcia-Molina, J. Hammer, K. Ireland, Y. Papakonstantinou, and J. Widom, "The TSIMMIS Project: Integration of Heterogeneous Information Sources," in 16th Meeting of the Information Processing Society of Japan, 1994.
17. L. Liu, C. Pu, and W. Han, "XWRAP: An XML-enabled Wrapper Construction System for Web Information Sources," in Proc. International Conference on Data Engineering (ICDE). San Diego, California, 2000.
18. M. L. Barja, T. Bratvold, J. Myllymaki, and G. Sonnenberger, "Informia: A Mediator for Integrated Access to Heterogeneous Information Sources," in Proceedings of the 1998 ACM CIKM International Conference on Information and Knowledge Management, G. Gardarin, J. C. French, N. Pissinou, K. Makki, and L. Bouganim, Eds. Bethesda, Maryland, USA: ACM Press, 1998, pp. 234-241.

19. C. R. Anderson, P. Domingos, and D. S. Weld, "Personalizing web sites for mobile users," in Proceedings of the 10th international conference on World Wide Web. Hong Kong: ACM Press, 2001, pp. 565-575.
20. G. A. Akerlof, "The Market for 'Lemons': Quality Uncertainty and the Market Mechanism," Quarterly Journal of Economics, vol. 84, pp. 488-500, 1970.
21. C. Shapiro and H. R. Varian, Information rules: a strategic guide to the network economy. Boston, Massachusetts, USA: Harvard Business School Press, 1999