A Roadmap Towards Distributed Web Assessment

Arno Scharl

University of Western Australia, Business School
35 Stirling Highway, Crawley, WA 6009, Australia
arno.scharl@uwa.edu.au

Abstract. The webLyzard project generates empirical Web data by processing large samples of Web sites automatically. It mirrors more than 5,000 international Web sites in monthly intervals and has amassed Web data in excess of one terabyte since 1999. Structural and textual analyses convert the wealth of information contained in the sample into detailed site profiles and aggregated content representations. A distributed approach promises to increase both sample size and the frequency of data gathering. This paper presents a roadmap towards distributed Web assessment, extending and revising the current system architecture to enhance its scalability and flexibility for investigating the dynamics of electronic content.

1 Introduction and Methodology

Software agents extracting and processing Web sites automatically ensure scalability, speed, consistency, rigorous structure, and abundant longitudinal data. They alleviate methodological limitations of subjective impressions and anecdotal evidence [1-3]. Judged against human evaluation, automated approaches are more efficient at handling dynamic Web data, and immune to inter- and intra-personal variances. These advantages come at the expense of sacrificing recipient-dependent attributes, which are difficult to quantify. Domain knowledge and expert opinions, therefore, play an important role when interpreting and applying the results.

This paper presents a roadmap towards distributed Web monitoring in terms of analytical objectives, process and data structures, and interface representation. Higher demands on transparency, flexibility and portability require new structures for data representation and transformation based on XML schemas and XSLT stylesheets, respectively. These data structures will pave the way for advanced visualisations. Topographic information mapping and geo-referenced projections will encourage the interactive exploration of Web resources.

2 Methodology

The webLyzard project has built the foundation for regularly sampling thousands of Web sites [2, 3]. As the volume and constantly changing char- acter of Web data entails ongoing analysis, a crawling agent mirrors the Web sites in monthly intervals, aptures their characteristics and stores the resulting profiles in a relational database.

N. Koch, P. Fraternali, and M. Wirsing (Eds.): ICWE 2004, LNCS 3140, pp. 171–175, 2004.

The move from centralised to distributed data gathering will not only increase the frequency of data gathering (weekly or daily intervals, depending on the site's characteristics), but also put samples of 100,000 sites and more within reach by leveraging spare bandwidth and previously unused computing resources from geographically dispersed clients.

The methodology considers both visible (raw text including headings, menus, or link descriptions) and invisible text (embedded markup tags, scripting elements, and so forth). Ignoring graphics and multimedia files, the agent follows a site's hierarchical structure until reaching 10 megabytes for regular sites, or 50 megabytes for online media. The size restriction helps compare systems of heterogeneous size, and manage storage capacity. Documents found in lower hierarchical levels are not part of the main user interface and can be disregarded for most analytical objectives.

2.1 Sample specification includes adding new Web sites, updating external rankings, and regularly checking the validity of addresses. The current sample of more than 5,000 Web sites comprises the Fortune 1000 (www.fortune.com), international media sites, environmental non-profit and advocacy organisations, European tourism sites, and nominees of the 2003 Webby Awards.

2.2 Data extraction processes and automatically codes the mirrored data. Automated coding attenuates subjective interpretations and many of the questionable aspects of manual coding. It renders the problems of time-consuming and expensive coder training or intra- and intercoder reliability obsolete, as the process measures variables directly and does not involve individuals who could disagree on particular attribute values. The variables constituting the site profiles fall into one of four categories: *navigational mechanisms, interactive features*, *layout and multimedia characteristics,* and *linguistic descriptives*.

2.3 Data representation of the current system needs a fundamental revision in order to meet the requirements of distributed Web monitoring (see Section 2.8). Currently, the variables are stored in a relational database and exported as comma-separated text files for further processing by external applications. The revision aims at replacing the current structures with a modular and reusable repository of Web metrics based on XML technologies. The flexibility of data definitions via XML schemas will provide the foundation for the envisioned distributed architecture. XSLT stylesheets will transform structural site metrics into diverse output formats such as XML, X(HTML), .CSV, and .PDF. The portability of XML-encoded data will encourage collaborative development of analytical modules.

2.4 Structural analysis relies upon multivariate statistical techniques or supervised neural networks to determine success factors by correlating the site profiles (set of independent variables) with measures of online success such as network statistics, server log-files, and questionnaire data (dependent variables). Important aspects of the proposed research are evaluating the availability and suitability of external rankings, and utilising third-party data such as Google Page Ranks (www.google.com), Alexa traffic statistics (www.alexa.com), and organisational details from Internet registrars (www.internic.net). The increasing availability of Application Programming Interfaces (APIs) and specialised scripting languages such as the Compaq Web Language [4] and the HTML Extraction Language [5] should facilitate these tasks.

2.5 Textual analysis converts the raw data into adequate representations, assigns languages to each document, and eliminates ambiguities in order to provide linguistic site metrics and identify focal points of environmental Web coverage. Most methods for analysing textual Web data originate from corpus linguistics and textual statistics [6, 7]. The type token ratio, for example, indicates the richness of the vocabulary of a given text by dividing the number of distinct words (types) by the total number of words (tokens) encountered when processing the text file.

The study of multicultural Web samples requires explicit attention to the role of language. *Trigrams* (three letter sequences within a document) and *short words* (determiners, conjunctions or prepositions) are good clues for guessing a language. As the results of both methods are nearly identical for textual segments comprising more than ten words [8], the computationally less intensive short word technique has been chosen for the current prototype. The proposed research aims to evaluate algorithms for document classification and employ more advanced parsing of sentence structures to eliminate ambiguities concerning the syntactic function or semantic nature of words, and explore knowledge representations such as XML topic maps [9] to handle dynamic and often redundant content segments.

2.6 Topic detection and tracking identifies semantic clusters in continuous streams of raw data that correspond to previously unidentified events [10]. Web-based tracking of events and public opinion on environmental issues complements and enhances traditional questionnaire surveys. Presenting news items culled from approximately 4,500 sources worldwide, Google News (news.google.com) demonstrates that automatic content classification by computer algorithms without human intervention can produce viable results. Standardised document type definitions – e.g., the News Markup Language (www.newsml.org), which is used by major international content providers such as *Reuters* or *United Press International* – facilitate classifying and aggregating content from multiple sources.

2.7 Visualisations of environmental Web content incorporates knowledge from cartography and geo-informatics [11], allowing visual recognition of document similarity by spatial cues and increasing the accessibility of complex data sets. The project will evaluate visualisation techniques and automate the creation of perceptual maps, currently performed manually by post-processing the output of statistical standard software. Scalable Vector Graphics (SVG), a modularised language for describing two-dimensional vectorial graphics in XML syntax [12], will be used to generate on-the-fly representations of Web content including frequency, context, geographic distribution, and hierarchical position of characteristic terms.

2.8 Distributed Web monitoring applies emerging computing frameworks that have revolutionised the processing of complex information. As some simulations are beyond the reach of current supercomputers, *Seti@Home* (setiathome.berkeley.edu), *Climateprediction.net* or *LifeMapper.org* exploit spare computer cycles by breaking down the calculations into parallel units that are processed by networks of volunteers (Seti@home scans radio signals from space to search for extraterrestrial intelligence;

Climateprediction.net explores climate change caused by manmade atmospheric pollutants; Lifemapper maps ecological profiles based on records of the world's natural history museums). As participating in active research attracts a large number of users, these projects often scale to thousands or millions of nodes. With the support of nearly 4.8 million users, Seti@Home executes more than 60 trillion floating point operations per second (TeraFLOPS) as of January 2004. This surpasses the capacity of the world’s fastest supercomputer – i.e., the *Earth Simulator,* a Japanese climate modelling project generating a maximum of 36 TeraFLOPS (www.es.jamstec.go.jp).

There are two main approaches to the distributed processing of information. *Peer-to-Peer (P2P) computing* has been popularised by file sharing and the highly parallel computing applications described above. *Grid computing*, by contrast, serves smaller communities and emphasises large-scale systems with fixed or slowly changing node populations in environments of at least limited trust [13, 14].

Migrating from centralised to distributed system architectures is complex and labour-intensive. Therefore, the project will utilise a standard service layer to collaborate with other projects, streamline system implementation, and manage computing resources more effectively. Examples of P2P service layers are the *Berkeley Open Infrastructure for Network Computing* (boinc.berkeley.edu), *XtremWeb* (www.lri.fr/~fedak/XtremWeb), and *JXTA* (www.jxta.org). The *Open Grid Services Architecture* (www.globus.org/ogsa) is based on Web service technologies to provide a more service-oriented platform, allowing computational services providers to register within known registries that can be browsed and searched by clients [15]. This achieves a cleaner middleware layer for individual applications based on industry standards such as the *Web Services Description Language* (WSDL) and the *Universal Description Discovery and Integration* (UDDI) protocol (www.w3.org/TR/wsdl; www.uddi.org).

While the project will benefit from the scalability and fault tolerance of P2P computing, its data-intensity suggests exploring light-weight Grid frameworks – as exemplified by the *Knowledge Grid* [16], a knowledge extraction service on top of the *Globus* (www.globus.org) grid architecture.

Distributed Web crawling permits gathering data in weekly or daily intervals, depending upon the media’s dynamic characteristics. Leveraging spare bandwidth and previously unused computing resources from geographically dispersed clients, samples of 100,000 sites and more become feasible, limited only by the number of participating individuals and institutions. As this process consumes lots of bandwidth but demands less resources in terms of processing capacity, it complements computation-intensive projects such as Seti@home and Climateprediction.net.

3 Conclusion

The roadmap presented in this paper incorporates knowledge of multiple disciplines to better understand the determinants, structure and success factors of Web content. The project builds upon the technology of the webLyzard project, which analyses more than 5,000 international Web sites in monthly intervals. Distributed data gathering will allow running the crawling agent in weekly or daily intervals. Leveraging spare bandwidth and previously unused computing resources from geographically dispersed clients, samples of 100,000 sites and more will become feasible, limited only by the number of participating individuals and institutions.

The use of open standards for encoding both structural and textual site data will enable other disciplines and stakeholders to effectively leverage the gathered information, which will be available via the collaborative environment of the Web portal (www.ecomonitor.net). A combination of the content management platform's core components and customised modules will facilitate the exchange of information with the research community, international and regional media, and the general public. This ensures a broad and international base of active researchers who are likely to contribute computational resources to the data gathering process, access and use the provided Web services, and promote the project among their peers.

References

[1] Ivory, M. Y.: Automated Web Site Evaluation: Researchers' and Practitioners' Perspectives. Kluwer Academic Publishers, Dordrecht (2003)

[2] Bauer, C., Scharl, A.: Quantitative Evaluation of Web Site Content and Structure. Internet Research: Networking Applications and Policy 10 (2000) 31-43

[3] Scharl, A.: Evolutionary Web Development. Springer, London (2000)

[4] Compaq Web Language. http://www.research.compaq.com/SRC/WebL/.

[5] Sahuguet, A., Azavant, F.: Building Intelligent Web Applications using Lightweight Wrappers. Data & Knowledge Engineering 36 (2001) 283-316

[6] McEnery, T., Wilson, A.: Corpus Linguistics. Edinburgh University Press, Edinburgh (1996)

[7] Biber, D., Conrad, S., Reppen, R.: Corpus Linguistics - Investigating Language Structure and Use. Cambridge University Press, Cambridge (1998)

[8] Grefenstette, G.: Comparing Two Language Identification Schemes. In: Proc. 3rd International Conference on Statistical Analysis of Textual Data (JADT-95) (1995) 263-268

[9] Le Grand, B.: Topic Map Visualization. In: J. Park and S. Hunting, (eds.): XML Topic Maps - Creating and Using Topic Maps for the Web. Addison-Wesley (2003) 267-282

[10] Chang, G., Healey, M. J., McHugh, J. A. M., Wang, J. T. L.: Mining the World Wide Web - An Information Search Approach. Kluwer Academic Publishers, Norwell (2001)

[11] Dodge, M., Kitchin, R.: New Cartographies to Chart Cyberspace. GeoInformatics 5 (2002) 38-41

[12] Quint, A.: Scalable Vector Graphics. IEEE Multimedia 10 (2003) 99-102

[13] Milenkovic, M., Robinson, S. H., Knauerhase, R. C., Barkai, D., Garg, S., Tewari, V., Anderson, T. A., Bowman, M.: Toward Internet Distributed Computing. Computer 36 (2003) 38-46

[14] Foster, I., Iamnitchi, A.: On Death, Taxes, and the Convergence of Peer-to-Peer and Grid Computing. In: F. Kaashoek and I. Stoica, (eds.): Peer-to-Peer Systems II: Second International Workshop, IPTPS 2003 Berkeley, CA, USA. Springer (2003) 118-128

[15] Giannadakis, N., Rowe, A., Ghanem, M., Guo, Y.-k.: InfoGrid: Providing Information Integration for Knowledge Discovery. Information Sciences 155 (2003) 199-226

[16] Cannataro, M., Talia, D.: The Knowledge Grid. Communications of the ACM 46 (2003) 89-93