Web Intelligence: Analysis, Representation & Deployment

Maurice MULVENNA[1], Sarabjot S. ANAND[1], Alex BÜCHNER[1], Matthias BAUMGARTEN[2,] Rüdiger BÖHM[2] & Zoë NEUMAN[2]
[1]MINEit Software, 5a Edgewater Business Park, Belfast, BT3 9JQ, Northern Ireland
Tel: +44 28 90368875 Fax: +44 28 90366068 Email: info @ mineit.com
[2]NIKEL, University of Ulster, Newtownabbey, BT37 0QB Northern Ireland
Email: nikel @ ulst.ac.uk

Abstract. This paper describes the practical application of web intelligence for visitor behavioural analysis. The objective of the paper is to show how the research in data mining carried out in the authors' laboratory has been synthesised to address the requirements of web intelligence. Initially, Web Intelligence is defined, and the data components used are explained. The web mining components of web intelligence are then outlined. Our research goals and our previous work is described in the following sections, followed by the inter-linked goals for digital marketing using web intelligence. The main body of the paper describes the various data-mining based analytical processes employed for web intelligence. The section on closing the marketing circle describes how the output of web intelligence can form a firm foundation for personalisation activities. The final section draws conclusions on the application of web intelligence in the near future.

1 Web Intelligence

Web Intelligence is the application of business intelligence software and methods to Internet data. It is defined as an infrastructural architecture containing data warehousing and data mining technologies, which draw upon a broad spectrum of components. These components are data warehousing staging and sorting tools, OLAP reporting and query tools, and data mining algorithms. In the present competitive environment, e-commerce organisations need to win and retain high-value customers to remain competitive. One technique that can be used to achieve greater loyalty from customers in Internet based retailing is to offer services that are predicated upon closing the marketing circle. These services include personalisation, recommendation, optimisation, etc. However for these services to succeed, "a library of rich visitor profiles must be present" [1]. Web intelligence is the means by which e-commerce companies can build these rich visitor profiles, utilising all the data that is generated by visitors and buyers at Internet e-commerce sites.

2 Web Data

The data available in electronic commerce environments is three-fold and includes server data in the form of log files, site specific web meta data representing the structure of the web site, and marketing information, which depends on the products and services provided [2,3]. Server data is generated by the interactions between the persons browsing an individual site and the web server. This data can be divided into log files and query data.

Historically, web servers recording server activity, errors and referrer information used a log file to record each event. It is now the standard that web servers use a combined log file

format, called Common Logfile Format [4]. This format combines the server and error logs into one file. More recently, the Extended Logfile Format [5] has been used, which consolidates the Common format with additional information, namely the referrer and cookie information. By incorporating referrer information, the output of the mining of these logfiles being much more useful and actionable in marketing terms. Cookies are tokens generated by the web server and held by the clients. The information stored in a *cookie* helps to ameliorate the transactionless state of web server http interactions, enabling servers to track client access across their hosted web pages. The logged cookie data is customisable and can contain keys for relating the navigational data to the content of the marketing data, including transactional data. Usually the following information is contained in a cookie: User ID, source IP address, time-to-live, randomly generated unique ID and user defined information.

A fourth data source that is typically generated on electronic commerce sites is *query data* to a web server. This data is usually generated when users of the web site use search or product locator facilities on the web site to search for relevant pages/products. This is often user interaction with a product database, via the company's Internet site.

The final source of data is web meta-data. This data describes the structure of the web site and is usually generated dynamically and automatically after a site update. Web meta-data generally includes neighbour pages, leaf nodes and entry points. This information is usually implemented as a site-specific index table, which represents a labelled, directed graph. Meta-data also provides information whether a page has been created statically or dynamically and whether user interaction is required or not. In addition to the structure of a site, web meta-data can also contain information of more semantic nature, usually represented in XML [6].

3 Web Mining Components of Web Intelligence

In the context of web intelligence, web mining may be defined as the application of data mining techniques to Internet data. This definition is sometimes extended to include statistical, database optimisation, and artificial intelligence techniques. Web mining has been sub-divided into web structure, web usage, and web content mining [7,8]. Web structure mining is the application of data mining techniques to web site structures. In many cases this may be the entire web, and research in intelligent search engines and intelligent agents is described in many articles, e.g., [9]. In our research, we define web structure mining as the mining of Internet data, *together* with data about the structure of the site. This may be thought of as enriching the efficacy of the data mining process with domain knowledge. The application of domain knowledge is further discussed in the analytical process section.

Web usage mining is the application of data mining to Internet web server log file data, which is described in the earlier section on web data. Web usage mining forms the core of our research in web mining for web intelligence, and log files provide the foundation data for visitor analysis. This type of analysis of the visitors to a web site can be subdivided into technographic and psychographic analysis [10].

Technographic analysis focuses on what is known about the visitor's technical platform, i.e., operating system, browser, plug-ins, user language, cookie information. On its own, this information is not a rich source of discriminatory data for visitor profiling but in conjunction with the homogenous data sets available after extract, transform & load operations to data warehousing, it contributes significantly. Psychographic analysis is the examination of what we know about the behavioural patterns of web site visitors. This includes the routes taken by visitors through a site, the time spent on each page, route differences based on differing entry points to site, aggregated route behaviour, general click stream behaviour, etc. This is the information of most use to web marketers, and is

equivalent to marketing intelligence about where shoppers enter the store, where shoppers go in the store, where they leave the store, what they look at but don't buy, what they buy and how quickly, etc.

Web content mining is the application of data and text mining algorithms and techniques to the contents of web pages, usually written in HTML. At its simplest, this entails the extraction of text between HTML tags for headings and titles, or the extraction of the HTML Meta tag content. However, as no rigidly enforced standards exist for the content of these tags, it is only by the use of advanced text mining techniques that utilise context that meaningful content may be used. Even so, there is no guarantee that the content is correct. Our research is based upon XML and RDF-based [11] data schemas that help to ensure correctness and proper context.

4 Research & Technology Goals

The NIKEL laboratory at the University of Ulster has long carried out data mining research [12-14]. Our interest and research in the application of data mining techniques to Internet data has being on-going since 1996 [15,16]. This paper represents the synthesis of our research in this area, and our progress to transform this research in the highly technical area of web mining into web intelligence software for Internet marketing. Our research goal in the fundamental data and web mining technology area are almost complete. We have developed web intelligence-optimised technology for segmentation, sequence & click stream detection [17], visualisation, and visitor activity analysis. Our technology goals have been to take the research developed in the laboratory and combine it into a software product. This goal is now complete, and the Easyminer product is marketed through a spinout company, MINEit Software [18].

5 Digital Marketing Goals

Organisations have typically invested a lot of money into developing their web sites and web strategy. Now they are seeking to assess the return they are receiving on their investment. Most sites use hits and page views as measure of success of the web site. According to a recent report by Forrester, however, using hits and page views as a measure of site success is like evaluating a musical performance by its volume [1]. Clearly the answer to this problem lies elsewhere. The technique used to measure the success of a web site obviously depends on what the goal of setting up the web site is in the first place. Using page hits as a metric does not provide a measure of success for any of these goals. Traditional marketing metrics such as churn rates, retention rates and revenues must be used as metrics for web success just as they are used in measuring the health of a business that is not on-line.

A new breed of software tools are now being developed to provide organisations with the opportunity to discover knowledge from the data collected from customer web interaction. This information allows companies to achieve their goals of personalised services and one-to-one marketing. These tools are collectively referred to as Web Intelligence tools. The following section describes the analytical processes used in the application of our web intelligence software called Easyminer.

6 The Analytical Processes

6.1 Segmentation

A starting point for traditional marketing is the segmentation of the customer base into smaller, more manageable groups of customers that have similar interests with respect to their interaction with the business. In the context of a traditional retailer and in the absence

of more customer data, this generally implies customers who buy similar products. In the case of an e-retailer, the web logs provide a large source of additional information about customers at no additional cost. However, E-retailers are not limited to analysing sales data. They can also discover similarities between customers based on their navigational behaviour, for example, products they may have browsed but not necessarily bought.

Segmentation based on a small number of attributes can be carried out manually or using a database query language. However, segmentation based on navigation behaviour is carried out based on a large number of attributes (the number of pages on the web site). Easyminer provides a variant of the k-means algorithm that has been adapted for use in web mining. Currently, two kinds of clustering may be undertaken: session clustering based on pages visited and the average time spent on the pages. The user specifies the number of segments that are expected to be present within the log file data and the minimum number of sessions needed for a cluster to be assumed as valid. The resulting pie chart shows the number of sessions within each of the segments while the cluster graph displays the spatial orientation of the segments, where smaller clusters (in terms of their diameter) represent more homogeneous segments. That is, sessions within the cluster are very similar to one another. On the other hand, a segment with a large diameter represents segments with sessions that are not very similar.

One of the challenges faced in segmentation of web log data is the high dimensionality of the data. Concept hierarchies defined on the documents can be used to reduce the dimensionality of the data. XML documents provide easy access to well defined domain knowledge as set by the Dublin core [19].

6.2 Sequences & Click Streams

Navigation of a web site is temporal in nature. Therefore, one of the basic forms of knowledge that needs to be discovered from data collected in web logs is navigational sequence knowledge. This describes the most commonly tread pathways through the web site, where a pathway is defined as based on a threshold value of sessions that follow the pathway, referred to as support. Easyminer uses the M*i*das sequence discovery algorithm [20] to discover sequences. Two types of sequences may be found using Easyminer: Open *sequences* and *click streams*. A sequence is a list of web page accesses ordered by the time of access within a session or across sessions for a particular customer. An open sequence is not necessarily a contiguous navigation of the web site. This means that an open sequence of the form `<index.html, orderform.html>` does not imply that there is a direct link between the index.html page and the orderform.html page that was navigated by customers that support this sequence. Customers supporting this sequence may have taken distinct paths from index.html to orderform.html; however, none of the individual paths navigated by the customer have the required support value to be considered as interesting within their own right. A click stream is a special type of sequence where the pages accessed have contiguous navigation. Thus a click stream of the form `<index.html, orderform.html>` does imply that a direct link exists between the index.html and orderform.html page and that the customers navigated this link during a particular session.

Three kinds of domain knowledge can be used within the discovery of sequences. These are navigational templates, network topologies and concept hierarchies. Navigational templates are used to tailor the sequences discovered from the log file to the users needs. Using these templates, goal-driven navigation pattern discovery is possible through the specification of start, end, as well as middle pages for sequences that are of interest to the user. A typical start locator is the home page, a customer support page, or a URL providing information about a special marketing campaign. A typical end page is a purchase page or a page for requesting more information. The second type of taxonomical domain knowledge is that of *network* topologies, which is useful when the topology of web site or a sub-

network of a large site is of interest to the user for the discovery of sequences. This domain knowledge is used to include or exclude parts of a web site from analysis.

Network topology domain knowledge within Easyminer is specified through a site map that is constructed from the log file being analysed. Sub-networks can be selected using point and click. In general, a network can be represented as a set of navigational patterns. The reason for distinguishing these two types of domain knowledge is that navigational templates are goal dependent and may change with each execution of M*i*das. A network on the other hand, is based on the structure of the web site, which is is less likely to change with the same frequency. Finally, concept hierarchies may also be specified and used to reduce the granularity of the discovered sequences in a similar way as their use within segmentation.

Two methods for visualising sequences exist within Easyminer. The first method uses the site map and overlays the sequences allowing the user to see the sequences within the context of the web site. The alternative method is to use a sequence tree view.

6.3 Visitor Data

In tandem with the powerful analytical tools available in web intelligence, visitor data can provide valuable marketing knowledge on the interactions between browsers and an e-commerce site. Typically, conventional log-analysis tools are based upon analysis of web server activity.

However, Easyminer provides additional capabilities to these tools. Firstly, the visitor information in Easyminer can be customised using profiles, whereby visitors may be identified as individual IP or domain addresses, or profiles can be constructed that describe sets of visitor activity. For example visitors from *UK domain at weekends only*, or *all visitors from academic institutions* (ac.uk, edu.cn, edu, etc.). Using these profiles, visitor groupings can be examined. The second additional capability in Easyminer is the ability to 'drill-down' into graphs and charts on visitor activity. In marketing terms, this facilitates examination of visitor activity from both a macro and micro perspective. Visitor activity analysis in the form of loyalty (frequency returning), page interest (time spent on page sequences), and deviation from designated profile groups is also provided.

7 Closing the Marketing Circle

Generating models with predictive capabilities is an important objective of our implementation. One use of such predictive models – click streams, segments, etc – is to provide a workbench for marketing analysts. Another more powerful application is to assist in *closing the marketing circle*. That is, to generate model output from our software, which is in a form that is also machine readable, by the web servers that host e-commerce sites upon which the analyses and model building have been based. In this situation, the e-commerce sites can produce real-time, dynamically generated pages tailored to individual or group profiles.

The tailoring made possible by predictive models hosted on web servers is very different from the two other types of tailoring popular in e-commerce sites. The first of these is 'check-box' tailoring, where a visitor customises a personal page. For example, a visitor can specify the region for the weather forecast, or which news feeds are required, etc. The second form of tailoring is sometimes called collaborative filtering, where a visitor specifies a profile by form filling. The profile is then matched against 'like-minded' visitors, and individualised content is fed back to the visitor based upon the collective preferences of these like minds.

The most important type of web page tailoring is personalisation, which can operate even while the visitor remains anonymous. If the visitor is identified with a cookie, or as a returning customer, then more targeted tailoring can take place. Personalisation is the

provision to the individual user of individually tailored products or services or information relating to products or service. Using the output of our predictive models, an e-commerce site can personalise content to the user, based upon the degree of membership that the visitor has to particular profiles. This can be a tailored banner ad, personalised prices for identified valuable returning customers, etc. An "improvement of the site based on interactions with all visitors" [21] is another objective that can be addressed using the predictive model's output.

8 Representing and Deploying Minded Knowledge

In order for an e-commerce web server to utilise the marketing knowledge (the output of the models), the knowledge must be made available to the server in a machine-readable form, and the web server must be able to act upon the knowledge. These two problems are called *representation* and *deployment.*

The most suitable vehicle for the representation of marketing knowledge output from our software is based upon XML. An important feature of XML-based representation is that it offers flexibility. It can be used by a web server and also transformed into a form easily understood by people. The Predictive Modelling Mark-up Language (PMML) [22] is a subset of XML which is being developed by a consortium of data mining vendors, under the aegis of the Data Mining Group. Currently, PMML has representative Document Type Definitions (DTD) for regression and decision trees. Our current research is adding DTDs for sequences and click streams.

The execution of the marketing knowledge by the web server is heavily dependent upon the system architecture in place. All of the following impact upon the deployment: server type (Netscape, Apache, IIS, etc); dynamic content from back-end databases or static pages; secure access for visitors (SHTTP); (reverse/forward) proxy server set-up; front-end packet sniffing/monitoring; and load-balancing architecture.

In technical terms, what is required to close the marketing circle is the capture of the click request from a visitor *before* the content/action that has been requested is served back to that visitor. It is only by performing this capture operation that the web server can decide 1) the appropriate PMML model, 2) the correct action to perform, 3) the content to source, 4) the dynamic construction of the page (e.g., banner ads, recommendations, personalisation, etc).

The capture operation can occur as a packet sniffing operation before the visitor request 'hits' the web server. However, most secure e-commerce sites make this a complicated operation, and the work-around (proxies, etc.) can add considerably to the server workload. Ultimately, the can result in reduced performance of the web server, and ultimately visitor disenchantment. One method that ameliorates this problem, is to pass-through the visitor request to the back-end content database (aka reverse proxy configuration), and process all the steps outlined above on either the content database server, or a third personalisation server. In this way, the web server architecture is unstressed, and optimisation of the personalisation server can tune-up the resulting system.

9 Conclusions & Future Work

This paper has described the application of some key research findings in a software tool called Easyminer. While Easyminer employs a number of data mining techniques to provide the user with useful web intelligence, it hides the complexity of these algorithms from the users who would typically be marketers and not data miners. Easyminer provides the means to discover the knowledge required to achieve goals such as one-to-one marketing and personalisation of web services. It also provides a means for measuring the success of on-line business and assesses the return on investment for web strategies.

The research underlying this software tool is on going. The technology required to close the circle with marketing automation components, such as marketing knowledge representation and execution, is well-advanced, and shows promising early results. Future research areas will concentrate on expanding the capabilities of the software to manage digital market interactions through other digital 'touch-points', including WAP telephony, PDAs, and digital interactive TV. In addition, work will also concentrate on expanding web intelligence into multi-channel marketing analysis, encompassing call centres and digital devices.

10 Acknowledgements

Portions of this research have been funded through grant support from EU MIMIC (RTD No. 26749), EU CERENA (RTD Proposal No. IST-1999-10039), and UK/EPSRC NetMODEL (RTD No. GR/N02986).

11 References

[1] Eric Schmitt, Harley Manning, Yolanda Paul, Joyce Tong. Measuring Web Success, Forrester Report, November, 1999

[2] Büchner, A.G., Mulvenna, M.D. Discovering Internet Marketing Intelligence through Online Analytical Web Usage Mining, ACM SIGMOD Record, ISSN 0163-5808, 27(4): 54-61, 1998.

[3] Mulvenna, M.D., Norwood, M.T. & Büchner, A.G. Data-driven Marketing, Int'l Journal of Electronic Commerce and Business Media, 8(3):32-35, 1998.

[4] World Wide Web Consortium (W3C), Logging Control In W3C httpd, 1995. http://www.w3.org/Daemon/User/Config/Logging.html

[5] World Wide Web Consortium (W3C), Extended Log File Format, W3C Working Draft, 1996. http://www.w3.org/TR/WD-logfile

[6] World Wide Web Consortium, W3c, 1997, http://www.w3.org/TR/REC-xml-19980210

[7] Büchner, A.G., Mulvenna, M.D., Anand, S.S. & Hughes, J.G. An Internet-enabled Knowledge Discovery Process, Proc. 9th Int'l. Database Conf., pp. 13-27, 1999

[8] Cheung, D.W., Kao, B., Lee, J., "Discovering User Access Patterns on the World-Wide Web", Proceedings 1st Pacific-Asia Conference on Knowledge Discovery and Data Mining, 1997

[9] Devlin, M., Scott, T.M., Mulvenna, M.D., Defining the Challenges for Communicating Information Using Multiple Agents on the WWW, UK Academy for Information Systems Annual Conference (UKAIS99), 1999

[10] Forrester Report, November, 1998

[11] World Wide Web Consortium, W3c, 1999, Resource Description Framework (RDF) Schema Specification. http://www.w3.org/TR/PR-rdf-schema

[12] Anand, S.S. Büchner, A.G. Decision Support using Data Mining, Financial Times Pitman Publishers, ISBN 0-273-63269-8, 1998.

[13] Anand, S.S., Scotney, B.W., Tan, M.G., McClean, S.I., Bell, D.A., Hughes, J.G., Magill, I.C. Designing a Kernel for Data Mining, IEEE Expert, 12(2):65-74, 1997.

[14] Anand, S.S., Patrick, A.R., Hughes, J.G., Bell, D.A. A Data Mining Methodology for Cross Sales, Knowledge-based Systems Journal, 1(1), 1999

[15] Büchner, A.G., Mulvenna, M.D. Discovering Behavioural Patterns in Internet Log Files: Playing the Devil's Advocate, 12th Biennial Int'l Telecommunications Society Conf. (ITS-98), Stockholm, Sweden, 1998.

[16] Mulvenna, M.D., Büchner, A.G. Norwood, M.T., Grant, C. The 'Soft-Push': Mining Internet Data for Marketing Intelligence, Working Conf. Electronic Commerce in the Framework of Mediterranean Countries, Metsovo, Greece, pp. 333-349, 1997.

[17] Büchner, A.G., Baumgarten, M., Mulvenna, M.D., Anand, S.S. & Hughes, J.G. Discovering Marketing-driven Navigation Patterns, ACM Workshop on Web Usage Analysis and User Profiling (WebKDD'99), 1999.

[18] MINEit Software, 1999, www.mineit.com

[19] Dublin Core. Dublin Core Metadata Element Set. http://purl.org/DC/

[20] Baumgarten, M., Büchner, A.G., Anand, S.S., Mulvenna, M.D., Hughes, J.G., Navigation Pattern Discovery from Internet Data, In: Masand, B., Spiliopoulou, M. (eds.) Advances in Web Usage Analysis and User Profiling, Lecturer Notes in Computer Science, Springer-Verlag, 2000

[21] Perkowitz, M., Etzioni, O., "Adaptive Web Sites: An AI Challenge", IJCAI 1997, Tokyo

[22] Predictive Model Markup Language (PMML), 1999, www.dmg.org/public/techreports/pmml-1_0.html